statistical methods for vqtl mapping and mendelian randomization analysis with a time-vayring exposure

by

yING cAO, bs, ma, ms

APPROVED:

pENG WEI, phd

Michael D. SWARTZ, phd

SUJA S. RAJAN, phd

DEAN, THE UNIVERSITY OF TEXAS

SCHOOL OF PUBLIC HEALTH

Copyright

by

Ying Cao, BS, MA, MS, PhD

2015

DEDICATION

To my parents

and

To my husband

statistical methods for vqtl mapping and mendelian randomization analysis with a time-vayring exposure

by

YING CAO

BS, Fudan University, 2007

MA, Indiana University Bloomington, 2010

MS, University of Texas School of Public Health, 2012

Presented to the Faculty of The University of Texas  
School of Public Health  
in Partial Fulfillment  
of the Requirements  
for the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas

School of Public Health

Houston, Texas

August 2015

Acknowledgements

With great pleasure I would like to express my sincere gratitude to my dissertation advisor, Dr. Peng Wei for helping me identify the research topics in this dissertation, for his excellent guidance and valuable suggestions, and also for the financial support he provided for this dissertation work. It was great to work with him in the past three years and I really appreciate all his help and supports.

I am deeply grateful to Dr. Michael Swartz for offering me a research assistant position when I was a master student. I thank him for trusting me when I had no previous research experience in the statistics field. I also thank Dr. Swartz for being my Ph.D. breadth advisor and for his supports all the time.

I thank Dr. Suja Rajan for being my minor advisor to give suggestions on course selection and more importantly, for her valuable comments on instrumental variable analysis. I appreciate her time and assistance.

A special thanks goes to Dr. Philip J. Lupo for serving as the external reviewer for my dissertation proposal defense and my final dissertation defense. I thank him for his comments and his valuable time.

I would also like to express my sincere gratitude to Dr. Taylor Maxwell for introducing the vQTL concept to me, which was the focus of the first part of my dissertation. I thank him for offering me the great research opportunity in the first two years of my Ph.D. study. I enjoyed the discussion and fun conversation with him.

I own my deepest gratitude to my parents for their support, encouragement, and love all the time. To my husband, Yaqing Fan, thank you for being with me, for your understanding, your encouragement, your believing, and your love.

statistical methods for vqtl mapping and mendelian randomization analysis with a time-vayring exposure

Ying Cao, BS, MA, MS, PhD

The University of Texas

School of Public Health, 2015

Dissertation Chair: Peng Wei, PhD

Complex diseases are affected by genetic factors, environmental factors, and their interactions. Traditional genetic factor study for complex diseases focuses on identifying loci associated with mean heterogeneity of a phenotype. A new class of genetic loci that are associated with phenotype variance heterogeneity (vQTL) has been suggested as candidates for identifying gene-gene and gene-environment interactions. While several tests have been proposed to detect vQTL for unrelated individuals, there are no tests for related individuals, commonly seen in family-based genetic studies. The first part of this dissertation introduces a likelihood ratio test (LRT) for vQTL identification using a linear mixed model framework, adjusting for covariates and family relatedness. The LRT test statistic approximately follows -distributions for normally distributed quantitative traits. Parametric bootstrap based LRT was proposed for non-normally distributed quantitative traits. Simulation studies show that the family-based test controls Type I error and has good power. We demonstrate the utility and efficiency gains of the proposed method using the Framingham Heart Study (FHS) data to detect loci associated with body mass index (BMI) variability.

The second part of this dissertation introduces Mendelian randomization (MR) analysis of a time-varying exposure using functional data analysis techniques. MR analysis is a method to analyze the causal effect of an environmental exposure variable on an outcome variable from observational studies by using genetic variants as instrumental variables. Many exposures of interest are time-varying, for example, BMI. However, current MR studies only use a single measurement of a time-varying exposure variable given that longitudinal measurements have been collected in many cohort studies. One measurement cannot adequately capture information of a time-varying exposure variable. We propose to use the functional principal component analysis method to recover the underlying individual trajectories of the time-varying exposure variable from the sparsely and irregularly observed longitudinal data, and then conduct MR analysis using the recovered trajectories. We focused on statistical testing for a causal effect. Different MR analysis methods have been proposed for continuous outcome variables and binary outcome variables to analyze the recovered functional exposure data. Simulation studies show that the functional data analysis-based methods incorporating longitudinal data have substantial power gain as compared with standard MR analysis. We used the FHS data to demonstrate the promising performance of the new methods.

Table of Contents

List of Tables i

List of Figures ii

Chapter 1: Background 1

Literature Review 1

vQTL Mapping 2

Mendelian Randomization 5

Public Health Significance 11

Chapter 2: A Family-based Joint Test for Mean and Variance Heterogeneity for Quantitative Traits 12

Summary 13

Introduction 14

Methods 17

A Likelihood Ratio Test for Detecting Mean and Variance Heterogeneity in Family-based Samples 17

Parametric Bootstrap for Non-normally Distributed Quantitative Traits 20

Simulation Studies 22

Application to the Framingham Heart Study 24

Results 25

Simulation Studies 25

Analysis of the FHS Data 27

Discussion 28

References 33

Supplemental Information 44

Chapter 3: Functional Data Analysis Approaches to Mendelian Randomization Study with a Time-varying Exposure 46

Summary 47

Introduction 48

Methods 52

Application to the FHS Data 57

Simulation Study 61

Discussion 64

References 67

Supplementary Materials 75

Chapter 4: Mendelian Randomization Analysis of a Time-varying Exposure for Binary Disease Outcomes using Functional Data Analysis Methods 82

Abstract 83

Introduction 84

Material and Methods 88

Notation 88

MR analysis using baseline measurement of a time-varying exposure variable 89

Functional data analysis methods for MR analysis with a time-varying exposure variable 89

Simulation Studies 93

Application to the FHS data 96

Results 98

Simulation Studies 98

The FHS Data Analysis 99

Discussion 102

References 105

Supplemental Data 117

Chapter 5: Conclusion and Future Directions 124

References 128

List of Tables

Chapter 2

Table 1. Empirical Type I error/power of association tests for normally distributed traits 39

Table 2. Empirical Type I error/power of association tests for t-distributed and -distributed quantitative traits. 41

Table 3. Summary of Framingham Heart Study BMI association analysis results. 43

Chapter 3

Table 1. P-values of testing the effect of longitudinal BMI on fasting glucose level using the FHS data 71

Table 2. Empirical Type I error rates of different methods 71

Table 3. Empirical power of different methods. 72

Chapter 4

Table 1. Empirical Type I error rates of different analysis methods in simulation set-up I. 110

Table 2. Empirical statistical power of different MR analysis methods in simulation set-up I. 110

Table 3. Empirical Type I error rates and statistical power of MR analysis methods in simulation set-up II. 111

Table 4. Analysis of the causal effect of BMI on the risk of T2D and CHD, and the effect of HDL on the risk of CHD using the FHS data. 111

List of Figures

Chapter 1

Figure 1. DAG for a Mendelian randomization analysis. 6

Chapter 2

Figure 1. BMI residual box plots and density plots by genotype for rs3120625 and rs7987059. 38

Chapter 3

Figure 1. DAG for a MR study 73

Figure 2. Longitudinal BMI observations from the FHS. 74

Chapter 4

Figure 1. Directed acyclic graph of MR analysis assumptions. 112

Figure 2. Directed acyclic graph of MR analysis with a time-varying exposure variable. 113

Figure 3. The observed FHS longitudinal data 114

Figure 4. PACE-predicted vs. observed BMI data 115

Figure 5. The estimated functional coefficient of the BMI GRS 116

# Chapter 1: Background

## Literature Review

Complex diseases are influenced by multiple genetic factors, behavioral and environmental factors, and their interactions. Most diseases that have a considerable burden on population health are complex diseases, such as obesity, type 2 diabetes, and others ([Wei et al., 2014](#_ENREF_36)).

In the past ten years, tremendous progress has been made in genetic studies for complex diseases. The technology advance of cost-effective and high-throughput single nucleotide polymorphism (SNP) array has enabled genome wide association (GWA) studies. GWA studies for complex diseases have identified many loci affecting the phenotype, while each of the loci has small effect and the sum of identified genetic variants only explains a portion of estimated heritability, less than half for most complex diseases that have been studied ([Stranger et al., 2011](#_ENREF_28)). Many hypotheses have been proposed to explain the missing heritability, including rare variants, gene-gene and gene-environment interactions ([Zuk et al., 2012](#_ENREF_41)). With the development of low-cost and high-throughput next generation sequencing (NGS) technology, rare variant association study for many complex traits are underway and various statistical models have been proposed ([Lee et al., 2014](#_ENREF_18)). In contrast, the identification of significant gene-gene and gene-environment interactions remains statistically challenging, majorly because of the curse of dimensionality and lack of statistical power. Variance-heterogeneity quantitative trait locus (vQTL), defined as the genetic locus associated with phenotype variability ([Rönnegård and Valdar, 2011](#_ENREF_23)), has been suggested for prioritizing the test for gene-gene and gene-environment interactions, thus reducing the dimension of total tests and aiding in identifying statistically significant interactions ([Paré et al., 2010](#_ENREF_21); [Struchalin et al., 2010](#_ENREF_30)). The statistical models that have been proposed for vQTL identification and the remaining challenges will be reviewed in detail in the vQTL mapping section.

In addition to genetic factors, the investigation of behavioral and environmental factors on complex traits has been the focus of conventional observational epidemiology studies. Although great success has been made, for example, the study of cigarette smoking on lung cancer risk ([Doll et al., 2004](#_ENREF_9); [Doll et al., 2005](#_ENREF_10)), many observational epidemiology studies that show significant change in disease risk cannot be confirmed by following randomized clinical trials (RCTs), for example, the association between Vitamin E intake and coronary heart disease (CHD) risk identified in observational study cannot be confirmed by RCT ([Hooper et al., 2001](#_ENREF_12)). The non-causal associations observed are mainly due to unadjusted confounders ([Vandenbroucke, 2004](#_ENREF_33)). Given the achieved success in genetic studies, would it be helpful for studying environmental exposure on complex diseases in observational studies? Mendelian randomization, a principle originally due to ([Katan, 1986](#_ENREF_15)), is to study the exposure-outcome relationship using genetic factors that affect the outcome only through the exposure. The progresses, limitations, and challenges in Mendelian randomization studies will be reviewed in the Mendelian randomization section.

### vQTL Mapping

Genetic studies of quantitative traits have been focusing on detecting loci associated with mean difference across genotypes ([Liu et al., 2013](#_ENREF_19); [Sun 2012](#_ENREF_31)). Most of the models used for detecting loci associated with mean difference assume equal variance of different genotypes. However, there has been increasing evidence for genetic loci associated with variability of quantitative traits and these loci are vQTLs ([Jimenez-Gomez et al., 2011](#_ENREF_13); [Shen et al., 2012](#_ENREF_24)). Phenotype variance heterogeneity across genotypes can be the result of an interaction effect with another genetic locus or environmental factor, known as gene-gene or gene-environment interactions ([Paré et al., 2010](#_ENREF_21); [Struchalin et al., 2010](#_ENREF_30)). If a locus interacts with another factor, the phenotype distribution of this genotype group will be a mixture of distributions with different means, leading to phenotype variance heterogeneity across genotype groups at this locus. Therefore, vQTL can serve as candidates for gene-gene and gene-environment interaction analyses, which will significantly reduce the dimension of interaction tests compared with pair-wise interaction screens. Moreover, vQTL can also be induced by linkage disequilibrium (LD) with a functional mean effect locus with a disparate minor allele frequency (MAF) ([Balding, 2009](#_ENREF_2); [Cao et al., 2014](#_ENREF_4)).

In light of the great promise of vQTL, different models and tests have been suggested for vQTL detection, for example, double generalized linear model (DGLM) ([Rönnegård and Valdar, 2011](#_ENREF_23)), squared residual value linear modeling (SVLM) ([Struchalin et al., 2012](#_ENREF_29)), and Levene’s test ([Paré et al., 2010](#_ENREF_21); [Shen et al., 2012](#_ENREF_24); [Struchalin et al., 2010](#_ENREF_30)). DGLM models the quantitative trait using a linear model and the square of residuals using a log linear model. The two models are optimized simultaneously. DGLM can incorporate adjustment for covariates, but is not robust to non-normality of quantitative traits. SVLM is similar to DGLM, by fitting the model for variance after getting the residuals from the linear model for mean. Levene’s test is an ANOVA F-test for the absolute difference between each observation and the median of its genotype group. Levene’s test is robust to non-normality, but cannot directly accommodate covariates. We recently proposed a likelihood ratio test (LRT) for vQTL detection in a linear model framework, which can adjust for covariates and control Type I error satisfactorily for non-normally distributed quantitative traits using parametric bootstrap ([Cao et al., 2014](#_ENREF_4)). The linear model based framework is versatile for a joint test for both mean and variance heterogeneity or a variance effect only test. Variance heterogeneity is known for being more difficult to detect than mean heterogeneity. A joint test can reveal more genetic loci associated with variance heterogeneity, which are also associated with mean heterogeneity, but may be missed by a variance effect only test ([Cao et al., 2014](#_ENREF_4); [Shen et al., 2012](#_ENREF_24)).

Family-based designs, such as the Framingham Heart Study (FHS) ([Splansky et al., 2007](#_ENREF_27)), have been widely used in genetic studies of complex traits. When compared with population-based designs, family-based designs have advantages that they are robust against population substructure, and allow tests for both linkage and association ([Laird and Lange, 2006](#_ENREF_16)). For quantitative traits, linear mixed effect models (LMMs) have been commonly used to account for familial correlations. Specifically, LMM models family relatedness as a random effect following a multivariate normal distribution with a zero mean vector and a covariance matrix proportional to the kinship matrix ([Aulchenko et al., 2007](#_ENREF_1); [Chen et al., 2013](#_ENREF_6); [Kang et al., 2010](#_ENREF_14); [Oualkacha et al., 2013](#_ENREF_20); [Yu et al., 2006](#_ENREF_40)). The kinship matrix represents the expected correlation for each pair of subjects that can be calculated based on pedigree information.

To the best of our knowledge, no vQTL test is yet developed specifically for family-based designs. To fill in this gap, we developed family-based LRTs (famLRTs) using a LMM framework for vQTL detection using family-based data. Chapter 2 includes the manuscript on the famLRTs.

### Mendelian Randomization

A Mendelian randomization study stands for an observational epidemiology study that intends to make causal inference of an exposure of interest on disease outcome using genetic variants that only affect the disease outcome through the exposure of interest ([Lawlor et al., 2008](#_ENREF_17)). Based on the law of independent assortment of Mendel, the inheritance of two different traits is independent, which generally holds with exclusion of LD. The transmission of alleles from parents to their children are independent, which says that alleles are randomly allocated during conception that the study of genetic variants on exposure variables is not subject to confounders ([Lawlor et al., 2008](#_ENREF_17)). At the population level, the associations between genetic factors and exposure variables are generally not confounded, in particular, not confounded by socioeconomic status and behavioral factors ([Smith et al., 2007](#_ENREF_26)). In addition, the association between genetic factors and disease risk cannot be due to reverse causality ([Lawlor et al., 2008](#_ENREF_17)). Therefore, genetic variants have been used as a proxy for an exposure to study its effect on disease outcome.

Mendelian randomization study is an application of instrumental variable (IV) analysis, which is a commonly used method in econometrics to deal with endogeneity. In a regression model, endogeneity occurs when an independent variable is correlated with the error term. For example, unobserved confounders can be a cause of endogenetiy ([Wooldridge, 2012](#_ENREF_37)). As depicted by the directed acyclic graph (DAG) in Figure 1, in a Mendelian randomization analysis, a genetic variant (G) is the IV that is used to study the effect of the exposure of interest (X) on the disease outcome (Y). The association between X and Y is confounded by some unmeasured confounders (U). To be a valid IV, the genetic variant must satisfy three assumptions: 1. the genetic variant is associated with the exposure variable; 2. the genetic variant is independent of the confounders that confound the association between the exposure and the disease outcome; 3. the genetic variant is independent of the disease outcome given the exposure and confounders. The third assumption is also known as the exclusion restriction, meaning that the genetic variant only affects the disease outcome through the exposure of interest. These three assumptions are sufficient for testing the association between the exposure and the disease outcome, while at least another assumption is needed for estimating the causal effect of the exposure of interest on the disease outcome, which is that all the associations in Figure 1 are linear and are not subject to interactions ([Lawlor et al., 2008](#_ENREF_17)).

**G**

**X**

**Y**

**U**

Figure 1. DAG for a Mendelian randomization analysis. G is a genetic variant (an instrumental variable (IV)), X is the exposure of interest, Y is the disease outcome, and U represents unmeasured confounders.

The most commonly used statistical method for IV analysis with a continuous outcome is two-stage lease square (2SLS). Specifically, in the first stage, an ordinary least square is performed for the exposure of interest using the IV(s) and measured covariates as the independent variables. In the second stage, another ordinary least square is performed for the outcome variable on the fitted values of the exposure of interest from the first stage and measured covariates. The fitted values of the exposure of interest using IV(s) are independent of unmeasured confounders, thus are independent of the error terms in the second stage linear model. Assuming that the relationship between the exposure of interest and the outcome is linear, the 2SLS estimator is consistent for estimating the causal effect of the exposure of interest on the outcome ([Wooldridge, 2010](#_ENREF_38)). A Wald test using heteroscedasticity-consistent standard error is often used for testing the significance of the causal effect.

The disease outcomes in epidemiology studies are often binary. IV analysis with binary outcome is more challenging. A straightforward extension from 2SLS is fitting a logistic or probit regression model in the second stage by replacing the exposure of interest with the fitted value from the first stage, which is often referred as two-stage predictor substitution (2SPS). However, due to a nonlinear relationship between the exposure of interest and the disease outcome, the 2SPS estimator is generally not consistent ([Terza et al., 2008](#_ENREF_32)). Instead of replacing the exposure variable with its fitted values, a second stage regression on the exposure variable and the residuals from the first stage is recommended for binary outcomes, which is known as two-stage residual inclusion (2SRI). The 2SRI estimator of the exposure effect on disease outcome is generally consistent ([Terza et al., 2008](#_ENREF_32)). The general idea is using the residuals from the first step regression as a proxy of the unmeasured confounders. 2SRI is also known as a control function approach for IV analysis ([Wooldridge, 2010](#_ENREF_38)) and has been used in Mendelian randomization studies with binary outcome ([Holmes et al., 2014](#_ENREF_11)).

Mendelian randomization analysis is a very attractive approach for causal inference analysis from observational studies, especially for the exposures of interest that cannot be studied using randomized clinical trials, such as body mass index (BMI), blood pressure, and high-density lipoprotein cholesterol (HDL-C). The success of large-scale GWA studies have provided many SNPs for studying the causal effects of exposures on complex diseases using Mendelian randomization analysis, for example, BMI associated SNPs have been used for studying the causal effect of BMI on CHD ([Holmes et al., 2014](#_ENREF_11)). The first assumption of using genetic variants as IV is often satisfied by careful selection of SNPs identified by successful GWA studies, and genetic variants are not associated with the unmeasured socioeconomic status and behavioral factors that may confound the association between the exposure of interest and disease outcome; however, there are possible violations of the exclusion restriction assumption of using genetic variants as IV, which are unique to genetic studies. Pleiotropy, meaning a genetic variant affects multiple traits, may cause bias in Mendelian randomization studies ([Lawlor et al., 2008](#_ENREF_17)). A good example is the variant of *APOE* gene. The variant affects many different exposures that are associated with the risk of myocardial infarction risk, including HDL-C, low-density lipoprotein cholesterol (LDL-C), postprandial lipaemia, and others ([Smith and Ebrahim, 2003](#_ENREF_25)). Therefore, *APOE* genotype is not a valid IV for studying the effects of any associated exposure on myocardial infarction risk ([Lawlor et al., 2008](#_ENREF_17); [Smith and Ebrahim, 2003](#_ENREF_25); [VanderWeele et al., 2014](#_ENREF_34)). LD can be another reason for violation of the exclusion restriction assumption ([Lawlor et al., 2008](#_ENREF_17); [VanderWeele et al., 2014](#_ENREF_34)). Specifically, if another genetic variant affects the disease outcome via other pathways and is in LD with the genetic variant used as an IV, it leads to association between the IV and the disease outcome in addition to the pathway through the exposure of interest, thus violating the IV assumption. In addition, population stratification, which occurs when different population subgroups have different allele frequency and different disease prevalence, can cause spurious association between genetic variants and disease outcome ([Cardon and Palmer, 2003](#_ENREF_5)). The risk of complex diseases is affected by a complicated biological system with contributions from many genetic and environmental factors. Therefore, there are other possible violations of using genetic variants as IV in Mendelian randomization studies ([Lawlor et al., 2008](#_ENREF_17); [VanderWeele et al., 2014](#_ENREF_34)).

Weak instrument problem is often encountered in IV analysis. It describes the situation that the IVs are weakly associated with the endogenous variable, thus causes biased effect estimates when the exposure-outcome relationship is confounded ([Bound et al., 1995](#_ENREF_3)). When the *F*-test statistic of IVs on the exposure variable in the first stage linear regression is greater than 11, the weak IV bias is often negligible, while substantial bias is introduced when the *F*-test statistic is smaller than 4 ([Pierce et al., 2011](#_ENREF_22)). In Mendelian randomization studies, a single genetic variant, usually a SNP, only explains a small proportion of the total variation of the exposure, and thus may cause weak IV problem. Genetic score (GS), a weighted count of effect alleles of multiple SNPs based on their estimated effects on the exposure of interest from previous studies, has been used to alleviate the weak IV problem and increase power in Mendelian randomization studies ([Holmes et al., 2014](#_ENREF_11); [Pierce et al., 2011](#_ENREF_22)). Moreover, using multiple large effect SNPs has been suggested to increase the power of Mendelian randomization studies ([Pierce et al., 2011](#_ENREF_22)). Using multiple SNPs has the additional benefit of alleviating the pleiotropy, LD, and population stratification problems that violate the IV analysis assumptions ([Lawlor et al., 2008](#_ENREF_17)).

Time-varying exposure is another challenge in Mendelian randomization studies. Many continuous exposures of interest are time-varying and may have a cumulative effect on the disease outcome, for example, BMI. Current Mendelian randomization studies often use the measured exposure at one time point ([Holmes et al., 2014](#_ENREF_11); [Voight et al., 2012](#_ENREF_35)), which is not adequate in capturing the long-term time-varying information. Using a single measurement of a time-varying exposure variable could underestimate the relationship between the exposure variable and the outcome variable ([Davis et al., 1990](#_ENREF_8)). Longitudinal measurements have been collected for many cohort studies, such as the FHS ([Splansky et al., 2007](#_ENREF_27)). How should the longitudinal measurements of the exposure of interest be incorporated in Mendelian randomization studies, especially when the longitudinal measurements are sparse and irregular? We propose to use functional data analysis techniques for conducting Mendelian randomization analysis of a time-varying exposure. The proposed methods for a continuous outcome variable are described in Chapter 3 and the proposed methods for a binary disease outcome are described in Chapter 4.

## Public Health Significance

The majority of diseases are complex diseases and many of them have high incidence rate in the U.S. and worldwide, including CHD, hypertension, type 2 diabetes, Alzheimer’s diseases, and more ([Craig, 2008](#_ENREF_7)). Complex diseases have been a big burden in public health. The development of complex diseases involves many genetic factors, environmental and behavioral factors, and their interactions. Unraveling the causal factors of complex diseases has been the focus of public health research. Although genetic studies of complex diseases have achieved big success, answering the missing heritability remains a big challenge. Gene-gene and gene-environment interactions have been attributed to the missing heritability of complex diseases ([Wu et al., 2012](#_ENREF_39); [Zuk et al., 2012](#_ENREF_41)). The statistical model development for vQTL identification provides powerful tools for detecting significant gene-gene and gene-environment interactions. In addition to genetic factors, environmental exposures are also big contributors to complex diseases. Observational studies of exposures on disease outcome often have unmeasured confounders ([Vandenbroucke, 2004](#_ENREF_33)). Mendelian randomization is a useful approach for causal inference analysis of the effects of modifiable exposures on disease outcome using genetic variants ([Lawlor et al., 2008](#_ENREF_17)). Statistical method development for analyzing time-varying exposures in Mendelian randomization studies provides useful tools for identifying the effect of time-varying exposures on complex disease outcome. Overall, the statistical method developments in my dissertation provide useful tools for identifying both genetic and environmental factors that contribute to complex disease development, thus lead to better understanding of the underlying causes of complex diseases.

# Chapter 2: A Family-based Joint Test for Mean and Variance Heterogeneity for Quantitative Traits

**Title of Journal Article**

A Family-based Joint Test for Mean and Variance Heterogeneity for Quantitative Traits

**Name of Journal Accepted**

Annals of Human Genetics

## Summary

Traditional quantitative trait locus (QTL) analysis focuses on identifying loci associated with mean heterogeneity. Recent research has discovered loci associated with phenotype variance heterogeneity (vQTL), which is important in studying genetic association with complex traits, especially for identifying gene-gene and gene-environment interactions. While several tests have been proposed to detect vQTL for unrelated individuals, there are no tests for related individuals, commonly seen in family-based genetic studies. Here we introduce a likelihood ratio test (LRT) for identifying mean and variance heterogeneity simultaneously or either effect alone, adjusting for covariates and family relatedness using a linear mixed effect model approach. The LRT test statistic for normally distributed quantitative traits approximately follows -distributions. To correct for inflated Type I error for non-normally distributed quantitative traits, we propose a parametric bootstrap based LRT that removes the best linear unbiased prediction (BLUP) of family random effect. Simulation studies show that our family-based test controls Type I error and has good power, while Type I error inflation is observed when family relatedness is ignored. We demonstrate the utility and efficiency gains of the proposed method using data from the Framingham Heart Study to detect loci associated with body mass index (BMI) variability.

Key Words: variance heterogeneity, QTL, linear mixed model, family data, BLUP

## Introduction

Genetic studies of quantitative traits have focused on detecting loci associated with mean difference across genotypes ([Liu *et al.*, 2013](#_ENREF_12), [Sun, 2012](#_ENREF_21)). However, there has been increasing evidence for genetic loci influencing variability of quantitative traits ([Jimenez-Gomez *et al.*, 2011](#_ENREF_8), [Shen *et al.*, 2012](#_ENREF_17)). The genetic locus associated with phenotype variability has been referred to as a variance-heterogeneity quantitative trait locus vQTL ([Rönnegård & Valdar, 2011](#_ENREF_16)). In addition, phenotype variance heterogeneity across genotypes can also be the result of an interaction effect with another genetic locus or environmental factor, known as gene-gene or gene-environment interactions ([Paré *et al.*, 2010](#_ENREF_14), [Struchalin *et al.*, 2010](#_ENREF_20)). If a locus interacts with another factor, the phenotype distribution of this genotype group will be a mixture of distributions with different means, leading to phenotype variance heterogeneity across genotype groups at this locus. Thus, vQTL can serve as candidates for gene-gene and gene-environment interaction analyses, which will significantly reduce the dimension of interaction tests compared with pair-wise interaction screens. Moreover, vQTL can also be induced by linkage disequilibrium (LD) with a functional mean effect locus with a disparate minor allele frequency (MAF) ([Cao *et al.*, 2014](#_ENREF_5), [Balding, 2009](#_ENREF_2)).

In light of the great promise of vQTL, different models and tests have been suggested for vQTL detection, for example, double generalized linear model (DGLM) ([Rönnegård & Valdar, 2011](#_ENREF_16)), squared residual value linear modeling (SVLM) of Struchalin *et al.* (2012) and Levene’s test ([Paré *et al.*, 2010](#_ENREF_14), [Shen *et al.*, 2012](#_ENREF_17), [Struchalin *et al.*, 2010](#_ENREF_20)). DGLM can adjust for covariates, but is not robust to non-normality of quantitative traits. In contrast, Levene’s test for variance heterogeneity is robust to non-normality, but cannot directly accommodate covariates. We recently proposed a likelihood ratio test (LRT) for vQTL detection in a linear model framework, which can adjust for covariates and control Type I error satisfactorily for non-normally distributed quantitative traits using parametric bootstrap ([Cao *et al.*, 2014](#_ENREF_5)). The linear model based framework is versatile for a joint test for both mean and variance heterogeneity or a variance effect only test. Variance heterogeneity is known for being more difficult to detect than mean heterogeneity. A joint test can reveal more genetic loci associated with variance heterogeneity, which are also associated with mean heterogeneity, but may be missed by a variance effect only test ([Cao *et al.*, 2014](#_ENREF_5), [Shen *et al.*, 2012](#_ENREF_17)).

Family-based designs, such as the Framingham Heart Study (FHS) ([Splansky *et al.*, 2007](#_ENREF_19)), have been widely used in genetic studies of complex diseases. When compared with population-based designs, family-based designs have advantages that they are robust against population substructure if certain analyses, such as the transmission disequilibrium test (TDT), are performed, and allow tests for both linkage and association ([Laird & Lange, 2006](#_ENREF_10)). For quantitative traits, linear mixed effect models (LMMs) have been commonly used to account for familial correlations. Specifically, LMM models family relatedness as a random effect following a multivariate normal distribution with a zero mean vector and a covariance matrix proportional to the kinship matrix ([Aulchenko *et al.*, 2007](#_ENREF_1), [Chen *et al.*, 2013](#_ENREF_6), Fisher, 1918, [Kang *et al.*, 2010](#_ENREF_9), [Oualkacha *et al.*, 2013](#_ENREF_13), [Yu *et al.*, 2006](#_ENREF_24)). The kinship matrix represents the expected correlation for each pair of subjects that can be calculated based on pedigree information.

To the best of our knowledge, no vQTL test is yet developed specifically for family-based designs, although traits from related samples could be transformed to be independent using GRAMMAR ([Aulchenko *et al.*, 2007](#_ENREF_1)) or GRAMMAR+ transformation (Belonogova *et al.*, 2013) and existing vQTL tests for independent samples could be subsequently employed. To fill in this gap, here we propose family-based LRTs (famLRTs) in a LMM framework for vQTL detection using family-based data. The tests are denoted as famLRTMV, famLRTM, and famLRTV for joint effects, mean effect only, and variance effect only, respectively. We demonstrate that the proposed famLRT test statistic approximately follows a -distribution when the quantitative trait of interest is normally distributed. However, p-value calculation for non-normally distributed traits requires parametric bootstrapping after removing the best linear unbiased prediction (BLUP) of the family random effect. Simulation studies show that existing vQTL tests for unrelated individuals have inflated Type I errors for family data, while famLRTs control Type I error well and have good statistical power under various scenarios. We also use simulations to compare the proposed famLRTs with the two-step approach based on the GRAMMAR+ transformation. Finally, we demonstrate the utility of the proposed tests using the FHS data to detect single nucleotide polymorphisms (SNPs) associated with body mass index (BMI) variability and confirmed one previously identified vQTL of BMI.

## Methods

### A Likelihood Ratio Test for Detecting Mean and Variance Heterogeneity in Family-based Samples

Let , , denote the quantitative trait value of subject , and , and be indicator variables indicating subject is major allele homozygous, heterozygous and minor allele homozygous, respectively. We model via a LMM:

where are the covariates of subject , is the regression coefficients of covariates, and are genotypes’ fixed effects on the phenotypic mean, is the random effect capturing familial correlation, and is the residual. To model variance heterogeneity of different genotype groups, we let , where , and are the residual variances of major allele homozygote, heterozygote and minor allele homozygote, respectively. The random effect vector and the residual vector are assumed to follow multivariate normal distributions and are independent of each other: and , where is two times the kinship matrix of size calculated using pedigree information representing the expected correlation for each pair of subjects, is a diagonal matrix of size with the th element on the diagonal being .

The joint test for mean and variance heterogeneity in family samples amounts to testing , vs. at least one “=” does not hold in model (1). We propose to employ a likelihood ratio test to test the full model (1) versus the reduced model (2):

Under the assumptions of model (1), , where the th row of is , and Let .The log-likelihood of the LMM (1) is:

For a given , the values of and that maximize the log-likelihood function can be written as:

Plugging the expressions of and , the profile log-likelihood of the LMM (1) can be written as:

Direct maximization of the profile log-likelihood is slow and numerically unstable given that is a sparse block diagonal matrix, especially when the number of pedigrees is large. To overcome this issue, we propose to employ generalized Cholesky decomposition to decompose as , where is a lower triangular matrix and is a diagonal matrix. We use the “gchol” function in the R package “kinship” to perform the generalized Cholesky decomposition in each iteration of the likelihood function maximization. Then and can be solved as an ordinary least square problem with new outcome vector and new design matrix . Finally, the profile likelihood function is maximized with respect to using the “BFGS” method in the R function “optim”.

The joint likelihood ratio test statistic for mean and variance heterogeneity in family samples is , where and are the maximum log-likelihood of full model (1) and reduced model (2), respectively. When the total sample size is large, the test statistic approximately follows a -distribution with 4 degrees of freedom. Similarly, to test for mean heterogeneity only while adjusting for variance heterogeneity in family samples, we test the full model (1) against the reduced model (3):

The likelihood ratio test statistic approximately follows a -distribution with 2 degrees of freedom when is large, where is the maximum log-likelihood of model (3). To test for variance heterogeneity only while adjusting for mean heterogeneity in family samples, we test the full model (1) against the reduced model (4):

The likelihood ratio test statistic approximately follows a -distribution with 2 degrees of freedom when is large, where is the maximum log-likelihood of model (4).

### Parametric Bootstrap for Non-normally Distributed Quantitative Traits

Variance tests are known to be more sensitive to deviation from normality; see, for example, page 189 of ([Lehmann, 1999](#_ENREF_11)). We previously observed that the variance test (LRTV) and the joint test for mean and variance heterogeneity (LRTMV) had inflated Type I error in unrelated samples when the residual normality assumption was violated, and we proposed a parametric bootstrap procedure to control the Type I error rate ([Cao *et al.*, 2014](#_ENREF_5)). In simulation studies for family samples, we also observed Type I error inflation of famLRTMV and famLRTV when residuals were not normally distributed, while famLRTM could control Type I error well. We intended to find a resampling method to accommodate adjustment for both covariates and family relatedness simultaneously. However, due to the complication introduced by the random effect accounting for familial correlation, none of the resampling methods we tried that simultaneously adjust for covariates and family relatedness controlled Type I error rate satisfactorily, for example, permutation of residuals and wild bootstrap for models with clustered errors ([Cameron *et al.*, 2008](#_ENREF_4)). We therefore propose to perform parametric bootstrap as in ([Cao *et al.*, 2014](#_ENREF_5)) after removing the family random effect, as follows:

1. Fit the reduced model (2) and obtain BLUP of random effect , denoted as , where and are the maximum likelihood estimate (MLE) of and in model (2), respectively.

2. Calculate the new outcome vector with BLUP removed as .

3. With the new outcome vector , the samples can be treated as independent and the famLRTMV test is reduced to LRTMV test for independent samples for which parametric bootstrap can be carried out as:

3a. Fit the null model where , and the full model , where . Then calculate LRTMV test statistic as in ([Cao *et al.*, 2014](#_ENREF_5)).

3b. Obtain MLE and of the null model in 3a and calculate residuals

, for .

3c. Permute ’s to generate ’s and make , for .

3d. Calculate test statistic by replacing with (for ) as in 3a.

3e. Repeat 3c and 3d for B times and the parametric bootstrap famLRTMV p-value is .

The parametric bootstrap famLRTV p-value can be calculated similarly by fitting the null model where , to obtain LRTV test statistic in 3a and calculating in 3d. As to be shown in the simulation studies, the parametric bootstrap based famLRTMV test and famLRTV test after removing BLUP of family random effect can control the Type I error rate satisfactorily for non-normally distributed residuals. Noticeably, an attractive property of the BLUP in step 1 is that it does not require the random effect **δ** to be normally distributed ([Robinson, 1991](#_ENREF_15)).

### Simulation Studies

To make the simulation studies representative of real family studies, we used pedigree information of 150 randomly selected families from the FHS and two real genome wide association study (GWAS) SNPs, one with MAF of 0.44 and the other with MAF of 0.14. We also used sex and age at the first clinical visit. The data we used for simulation studies includes 1019 individuals. The largest pedigree has 20 individuals and the smallest pedigree has 3 individuals.

We first evaluated the performance of famLRTs for normally distributed quantitative traits in four different scenarios: 1. There is no association between the SNP and the simulated quantitative trait; 2. The SNP is only associated with the mean heterogeneity of the trait; 3. The SNP is only associated with the variance heterogeneity of the trait; 4. The SNP is associated with both mean and variance heterogeneity of the trait. The model , for , was used to simulate quantitative trait ’s, where and are indicator variables indicating heterozygote and minor allele homozygote, respectively; is the family random effect; and is the residual. was simulated from the multivariate normal distribution , where is two times the kinship matrix. In scenario 1 and 3 without mean effect, we let . In scenario 2 and 4 with mean effect, we let and .35. Similarly, in scenario 1 and 2 without variance effect, was generated randomly from . In scenario 3 and 4 with variance effect, was generated randomly from for the common allele homozygote, for the heterozygote, for the minor allele homozygote. For both of the GWAS SNPs, 1000 simulated datasets were generated for each of the four scenarios. Empirical Type I error/power of famLRTMV, famLRTM, famLRTV, parametric bootstrap based famLRTMV and famLRTV were calculated at significance level of 0.05 and 0.01. For parametric bootstrap based tests, 1000 resamplings were conducted for each simulated dataset. LRTs ignoring familial correlation, including LRTMV, LRTM, and LRTV, were included in scenario 1 to evaluate the impact on Type I error when family relatedness was ignored. Of note, the recently proposed GRAMMAR+ method transforms traits from related individuals to be independent and preserves high power in the subsequent independent sample test (Belonogova *et al.*, 2013). We therefore also included LRTMV, LRTM, and LRTV tests using GRARMMAR+ residuals for comparison.

For non-normally distributed quantitative traits, we conducted two sets of simulation studies, one with residuals generated from *t*-distributions representing quantitative traits with heavy tails, and the other with residuals generated from -distributions representing skewed quantitative traits. For each set of simulation studies, we also considered two SNPs with different MAFs in four scenarios and used the same model for quantitative trait generation as we did for normally distributed traits. The only difference was residual generation. For heavy tail quantitative traits, residuals were generated randomly from a *t*-distribution (df=10) in scenario 1 and 2, and from *t*-distributions with df=10, 5, and 3 for the common allele homozygote, heterozygote, and minor allele homozygote, respectively, in scenario 3 and 4. For skewed quantitative traits, residuals were generated randomly from a -distribution (df=3) in scenario 1 and 2, and from -distributions with df=3, 4, and 6 for the common allele homozygote, heterozygote, and minor allele homozygote, respectively, in scenario 3 and 4. Residuals from -distributions were scaled to have mean 0 in each genotype group and then divided by 3 to make variances comparable to variances in other simulation scenarios. We simulated 1000 replicates for each scenario. Empirical Type I error/power was calculated at significance levels of 0.05 and 0.01. For parametric bootstrap based famLRTMV and famLRTV, 1000 resamplings were conducted for each replicate to obtain parametric bootstrap p-values.

### Application to the Framingham Heart Study

The Framingham Heart Study (FHS) is a family-based study conducted for three generations: the Original Cohort, the Offspring Cohort, and the Third Generation Cohort ([Splansky *et al.*, 2007](#_ENREF_19)). We downloaded the FHS genotype and phenotype data from NCBI dbGaP (http://www.ncbi.nlm.nih.gov/gap). We examined the association between mean or variance heterogeneity of BMI and SNPs near previously identified SNPs associated with BMI. The phenotype information was extracted from the fourth clinic exam for the Original Cohort, the third clinic exam for the Offspring Cohort, and the first clinic exam for the Third Generation Cohort to make the average age when phenotype information was recorded comparable across the three cohorts. BMI was calculated as weight in kilograms divided by the square of height in meters. Large-scale meta-analyses have identified SNPs associated with mean heterogeneity of BMI ([Berndt *et al.*, 2013](#_ENREF_3), [Speliotes *et al.*, 2010](#_ENREF_18), [Willer *et al.*, 2009](#_ENREF_22)) and variance heterogeneity of BMI ([Yang *et al.*, 2012](#_ENREF_23)). We extracted SNPs within +/- 500 kilobase pairs (kb) of previously identified BMI associated SNPs from the genotype data of FHS SNP Health Association Resource (SHARe) GWAS. 6484 individuals from three generations with both phenotype and genotype information available were included in the analysis. We restricted our analysis to SNPs with at least 20 observations in each genotype group. 4640 SNPs were tested using famLRTM, and parametric bootstrap based famLRTMV and famLRTV, adjusting for sex, age, and age squared. 1,000,000 resamplings were performed for parametric bootstrap based tests.

## Results

### Simulation Studies

Table 1 shows the simulation study results for normally distributed quantitative traits. For family data, tests without adjustment for family relatedness (LRTMV, LRTM and LRTV) have inflated Type I errors for both SNPs with different MAFs, while tests adjusting for family relatedness, including parametric tests (famLRTMV, famLRTM and famLRTV), parametric bootstrap based tests (famLRTMV(PB) and famLRTV(PB)) and LRTs using GRAMMAR+ residuals, control Type I errors satisfactorily at significance levels of 0.05 and 0.01 based on 1000 replications (Table 1A). To assess powers of family-based tests, we conducted simulation studies in three scenarios (Table 1B). When a quantitative trait only shows mean difference across genotype groups, famLRTM is the most powerful test; famLRTMV loses power due to larger degrees of freedom; famLRTV controls Type I error well. Similarly, when there is only variance difference of a quantitative trait across different genotype groups, famLRTV is the most powerful test; famLRTMV also loses power in this scenario; famLRTM controls Type I error well. In contrast, when a SNP is associated with both mean and variance heterogeneity of a quantitative trait, the joint test is more powerful than either of the tests for a single effect. We noticed that powers of parametric bootstrap based famLRTMV and famLRTV are slightly lower compared with their parametric versions, but the power loss is moderate. The two-step tests using GRAMMAR+ residuals have power comparable to the one-step famLRTs in all three scenarios. In addition, with the same magnitude of simulated effects, tests for the SNP with lower MAF have lower power than tests for the SNP with higher MAF. This is as expected because sample sizes of heterozygotes and minor allele homozygotes decrease as MAF decreases.

Table 2 shows the simulation study results for non-normally distributed quantitative traits, one set simulated from *t*-distributions representing heavy tail traits and the other set simulated from -distributions representing skewed traits. Parametric tests for variance, including both famLRTMV and famLRTV, cannot control Type I errors when quantitative traits are not normally distributed, while the mean only parametric test famLRTM still performs satisfactorily. In contrast, parametric bootstrap based tests for variance can fix the problem and control Type I errors at the nominal levels for both heavy tail and skewed traits (Table 2A). We then evaluated the power of the tests that can control Type I errors in three different scenarios (Table 2B). The same as the simulation study results for normally distributed traits, the joint test is the most powerful when a SNP is associated with both trait mean and variance heterogeneity, while it sacrifices power in a mean effect or variance effect only scenario because of the extra degrees of freedom of the test. For the mean effect only scenario, parametric bootstrap famLRTV controls Type I errors well, while for the variance effect only scenario, famLRTM controls Type I errors well. The powers of all tests decrease as the MAF of the associated SNP decreases because sample sizes of heterozygotes and minor allele homozygotes decrease. We also applied the LRTs for unrelated individuals to the GRAMMAR+ residuals for simulated non-normally distributed traits. As expected, although the GRAMMAR+ transformation can effectively remove family relatedness, it does not overcome the issue of inflated Type I error for vQTL test (results not shown).

### Analysis of the FHS Data

We investigated the association between BMI and BMI candidate SNPs using the family-based FHS data. BMI candidate SNPs are SNPs within +/- 500kb of previously identified 50 SNPs significantly associated with either mean or variance heterogeneity of BMI by large-scale meta-analysis ([Berndt *et al.*, 2013](#_ENREF_3), [Speliotes *et al.*, 2010](#_ENREF_18), [Willer *et al.*, 2009](#_ENREF_22)). As shown in Supplemental Figure 1, the distributions of both BMI and BMI residuals are quite positively skewed and we would expect inflated Type I errors for famLRTMV and famLRTV. As a result, we applied the proposed parametric bootstrap-based famLRT tests. The effective number of tests of the 4640 BMI candidate SNPs is 2731, corresponding to explanation of 99.5% of the total variation of 4640 SNPs, which was calculated following the method proposed by ([Gao *et al.*, 2008](#_ENREF_7)). However, none of the tested SNPs passed the significance level of after multiple test correction. The top 10 SNPs with the smallest p-values of parametric bootstrap based famLRTMV are listed in Table 3. Four SNPs are in the same region near gene *CADM2*. Out of the ten SNPs, nine SNPs have p-values less than 0.05 for both the mean effect only and the variance effect only tests, showing evidence of heterogeneity in both mean and variance. For seven SNPs, the mean and variance joint test p-value is smaller than the p-values of both individual effect tests, confirming that the test for both mean and variance heterogeneity is more powerful for detecting vQTLs when a SNP is also associated with mean heterogeneity. BMI residual plots after removing covariate effects and family relatedness show observable variance heterogeneity across genotypes for the top two SNPs rs3120625 and rs7987059 (Figure 1). Furthermore, rs12328474 was identified as a top SNP associated with BMI variability by a large-scale genetic association study, with p-value of in a discovery set of 104,640 samples and p-value of 0.02 in a replication set of 32,403 samples ([Yang *et al.*, 2012](#_ENREF_23)). Our analysis confirmed the vQTL of BMI with parametric bootstrap based famLRTv p-value of using the 6484 FHS samples (Table 3). In contrast, if we performed vQTL test in 2380 unrelated FHS samples using the LRTV test, the p-value for rs12328474 is 0.22, which is much less significant than the famLRTV p-value based on all FHS samples. In addition, for the top ten SNPs listed in Table 3, all the tests using independent samples have less significant p-values than their corresponding family-based tests using all samples. This demonstrates the power gain enabled by our proposed family-based tests on all related individuals in a family design.

## Discussion

In this paper, we have proposed a joint test of mean and variance heterogeneity for family data in the LMM framework. The model is flexible in that it can accommodate a joint test and a mean or variance effect only test with adjustment for both covariates and family relatedness. To our knowledge, famLRTMV is the first test developed for jointly testing mean and variance effects for family data.

Family design is common in genetic studies. We have demonstrated that the joint test for mean and variance heterogeneity has inflated Type I error when familial correlation is ignored. An alternative is to extract unrelated individuals for analysis, which will significantly reduce the total sample size and lead to power loss. For example, 6484 individuals were included in our case study of the FHS data, while only 2380 individuals out of the 6484 individuals are unrelated. We would discard approximately two thirds of the entire sample if only unrelated samples were considered, which is a big loss of valuable phenotype and genotype information that has been collected.

We showed that the famLRTMV test statistic of a normally distributed trait approximately follows a -distribution, while Type I error inflation was observed for a non-normally distributed trait. Transformation of a non-normally distributed trait to be normal is often used in linear models, including Box-Cox transformation and inverse normal transformation. However, transformation is not preferred for a variance test because it may distort the variance related information in the untransformed trait, which may lead to bias and loss of power ([Cao *et al.*, 2014](#_ENREF_5)). Resampling based methods are more appropriate for p-value calculation in this scenario. Commonly used resampling methods for LMMs, including wild bootstrap ([Cameron *et al.*, 2008](#_ENREF_4)), case bootstrap and bootstrap for both random effects and residuals ([Cameron *et al.*, 2008](#_ENREF_4)), are not applicable to the model we proposed. Wild bootstrap only takes into account of mean heterogeneity, but not variance heterogeneity. Case bootstrap is not appropriate because different families may have very different sizes so that the total sample size of each resampled sample using case bootstrap may vary dramatically. Due to the fact that each family has a unique correlation structure and the correlation within each family is not exchangeable, bootstrap for random effects is not applicable either. Therefore, we suggested a two-step approach for parametric bootstrap by removing the BLUP of the family random effect at the first step and performing parametric bootstrap for the uncorrelated samples in the second step. Decreased power was observed for the two-step parametric bootstrap based famLRTMV when compared with one-step parametric famLRTMV (Table 1B). The power decrease is subtle for the variance effect only and joint effect scenarios, while it is a little more for the mean effect only scenario. However, the two-step approach is necessary for Type I error control when the quantitative trait of interest is not normally distributed. In addition, removing the BLUP of the family random effect avoids decomposing the sparse block diagonal kinship matrix when performing parametric bootstrap, which is much more computational efficient. For genome-wide pedigree-based QTL analysis, Aulchenko *et al*. [(2007](#_ENREF_1)) also suggested a two-step approach to shorten computation time by removing familial correlation in the first step. Furthermore, for a given quantitative trait of interest, the empirical distribution of the famLRTMV test statistic constructed based on parametric bootstrap is invariant to the SNP being tested after removal of the BLUP of the family random effect (Supplemental Figure 2). In other words, the parametric bootstrap based famLRTMV only needs to be conducted for one SNP given a trait and the empirical distribution of the test statistic can then be used to calculate p-values for all the SNPs. Although the resampling based method is necessary for famLRTMV p-value calculation when the quantitative trait is not normally distributed, the fact that only one set of resampling suffices for all the SNPs makes the test much more practical and useful.

To conduct famLRTMV for a non-normally distributed quantitative trait in a large scale association analysis, e.g., GWAS, we suggest an initial screening by calculating parametric p-values for all the SNPs based on -distribution. If there are any SNPs passing the significance level, one set of parametric bootstrap can then be conducted to construct the empirical null distribution of the test statistic, which can be used to determine the significance of all the SNPs (Supplemental Figure 2).

In addition to famLRTs, independent sample tests using GRAMMAR+ residuals can be an option for vQTL detection using family data as shown in Table 1. GRAMMAR+ transformation is an approximation method (Belonogova *et al.*, 2013) and any tests using GRAMMAR+ residuals are two-step tests where the covariate effects are removed in the first step. Although tests using GRAMMAR+ residuals have power comparable to famLRTs shown by our simulation study (Table 1), we would expect power loss in the mean test part if the tested SNP is correlated with any covariates ([Cao *et al.*, 2014](#_ENREF_5)). The advantage of GRAMMAR+ transformation is that GRAMMAR+ residuals can be tested using Levene’s test or Fligner-Killeen test for variance heterogeneity. These two tests are robust to non-normality and are computationally efficient. As for computational time, GRAMMAR+ transformation followed by LRTMV took approximately 90 minutes, while famLRTMV took approximately 5 minutes for 1000 replications of the simulated data with 1019 subjects on the same computer.

As demonstrated by simulation studies and application to the FHS data, the joint test for detecting mean and variance heterogeneity is more powerful than a single effect test when a SNP is associated with both mean and variance heterogeneity of a quantitative trait. Mean heterogeneity is a more common phenomenon than variance heterogeneity in QTL analysis ([Yang *et al.*, 2012](#_ENREF_23)). In addition, vQTLs can occur due to LD with a causal functional SNP with mean heterogeneity. In this case the vQTL will show mean heterogeneity as well ([Cao *et al.*, 2014](#_ENREF_5)). The real data analysis showed that nine out of the ten top SNPs are associated with both mean and variance heterogeneity of BMI using nominal significance level of 0.05 (Table 3). By both simulation (Table 1B and 2B) and empirical studies, we found that the power loss of the joint test is subtle compared with the variance only test when a SNP has variance only effect. For example, rs2889756 has a joint test p-value of , which is even smaller than the variance test p-value of given a mean test p-value of 0.12; rs12328474 has a joint test p-value of , which is only slightly larger than the variance test p-value of given a mean test p-value of 0.26. Therefore, we recommend vQTL screening using joint effect test famLRTMV, followed by variance effect only test famLRTV for the significant SNPs identified by famLRTMV test to increase the power of vQTL detection. The significance level of the subsequent famLRTV test only needs to be adjusted for the number of SNPs that are globally significant based on famLRTMV test. Our previous simulation study has demonstrated that a variance test after rejecting the null hypothesis of the joint test can control family-wise Type I error rate at the nominal level ([Cao *et al.*, 2014](#_ENREF_5)). The famLRTMV, famLRTM and famLRTV tests are implemented as R functions, which will be posted on our website at https://sites.google.com/site/utpengwei/

**Acknowledgement**

This work was supported by the National Institutes of Health grants R01HL105502 (to TJM), R01HL116720 and R01CA169122 (to PW). The authors declare that there are no conflicts of interest. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL-64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University. We thank the anonymous reviewers for their constructive comments that improved the presentation of the paper.

## References

Aulchenko, Y.S., De Koning, D.-J. & Haley, C. (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics,* 177**,** 577-585.

Balding, D.J. (2009) The advance of Bayesian methods for genetic association analysis *Presented at the 59th Annual Meeting of The American Society of Human Genetics.,* November 22, 2009, Honolulu, HI.

Belonogova, N. M., Svishcheva, G. R., van Duijn, C. M., Aulchenko, Y. S., & Axenovich, T. I. (2013). Region-based association analysis of human quantitative traits in related individuals. *PloS one*, 8(6), e65395.

Berndt, S.I., Gustafsson, S., Maegi, R., Ganna, A., Wheeler, E., Feitosa, M.F., Justice, A.E., Monda, K.L., Croteau-Chonka, D.C. & Day, F.R. (2013) Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature genetics,* 45**,** 501-512.

Cameron, A.C., Gelbach, J.B. & Miller, D.L. (2008) Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics,* 90**,** 414-427.

Cao, Y., Wei, P., Bailey, M., Kauwe, J.S. & Maxwell, T.J. (2014) A versatile omnibus test for detecting mean and variance heterogeneity. *Genetic Epidemiology,* 38**,** 51-59.

Chen, H., Meigs, J.B. & Dupuis, J. (2013) Sequence kernel association test for quantitative traits in family samples. *Genetic Epidemiology,* 37**,** 196-204.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the royal society of Edinburgh*, 52(2), 399-433.

Gao, X., Starmer, J. & Martin, E.R. (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology,* 32**,** 361-9.

Jimenez-Gomez, J.M., Corwin, J.A., Joseph, B., Maloof, J.N. & Kliebenstein, D.J. (2011) Genomic analysis of QTLs and genes altering natural variation in stochastic noise. *PLoS genetics,* 7**,** e1002295.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-Y., Freimer, N.B., Sabatti, C. & Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature genetics,* 42**,** 348-354.

Laird, N.M. & Lange, C. (2006) Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet,* 7**,** 385-94.

Lehmann, E.L. (1999) *Elements of large-sample theory.* New York: Springer.

Liu, C.T., Monda, K.L., Taylor, K.C., Lange, L., Demerath, E.W., Palmas, W., Wojczynski, M.K., Ellis, J.C., Vitolins, M.Z., Liu, S., Papanicolaou, G.J., Irvin, M.R., Xue, L., Griffin, P.J., Nalls, M.A., Adeyemo, A., Liu, J., Li, G., Ruiz-Narvaez, E.A., Chen, W.M., Chen, F., Henderson, B.E., Millikan, R.C., Ambrosone, C.B., Strom, S.S., Guo, X., Andrews, J.S., Sun, Y.V., Mosley, T.H., Yanek, L.R., Shriner, D., Haritunians, T., Rotter, J.I., Speliotes, E.K., Smith, M., Rosenberg, L., Mychaleckyj, J., Nayak, U., Spruill, I., Garvey, W.T., Pettaway, C., Nyante, S., Bandera, E.V., Britton, A.F., Zonderman, A.B., Rasmussen-Torvik, L.J., Chen, Y.D., Ding, J., Lohman, K., Kritchevsky, S.B., Zhao, W., Peyser, P.A., Kardia, S.L., Kabagambe, E., Broeckel, U., Chen, G., Zhou, J., Wassertheil-Smoller, S., Neuhouser, M.L., Rampersaud, E., Psaty, B., Kooperberg, C., Manson, J.E., Kuller, L.H., Ochs-Balcom, H.M., Johnson, K.C., Sucheston, L., Ordovas, J.M., Palmer, J.R., Haiman, C.A., Mcknight, B., Howard, B.V., Becker, D.M., Bielak, L.F., Liu, Y., Allison, M.A., Grant, S.F., Burke, G.L., Patel, S.R., Schreiner, P.J., Borecki, I.B., Evans, M.K., Taylor, H., Sale, M.M., Howard, V., Carlson, C.S., Rotimi, C.N., Cushman, M., Harris, T.B., Reiner, A.P., Cupples, L.A., North, K.E. & Fox, C.S. (2013) Genome-wide association of body fat distribution in African ancestry populations suggests new loci. *PLoS genetics,* 9**,** e1003681.

Oualkacha, K., Dastani, Z., Li, R., Cingolani, P.E., Spector, T.D., Hammond, C.J., Richards, J.B., Ciampi, A. & Greenwood, C.M. (2013) Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genetic Epidemiology,* 37**,** 366-376.

Paré, G., Cook, N.R., Ridker, P.M. & Chasman, D.I. (2010) On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS genetics,* 6**,** e1000981.

Robinson, G.K. (1991) That BLUP is a good thing: The estimation of random effects (with discussion). *Statist. Sci.,* 6.

Rönnegård, L. & Valdar, W. (2011) Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics,* 188**,** 435-447.

Shen, X., Pettersson, M., Rönnegård, L. & Carlborg, Ö. (2012) Inheritance beyond plain heritability: variance-controlling genes in Arabidopsis thaliana. *PLoS genetics,* 8**,** e1002839.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Allen, H.L., Lindgren, C.M., Luan, J.A. & Mägi, R. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics,* 42**,** 937-948.

Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'agostino, R.B., Fox, C.S., Larson, M.G. & Murabito, J.M. (2007) The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *American journal of epidemiology,* 165**,** 1328-1335.

Struchalin, M.V., Dehghan, A., Witteman, J.C.M., Van Duijn, C. & Aulchenko, Y.S. (2010) Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. *BMC genetics,* 11**,** 92.

Struchalin, M. V., Amin, N., Eilers, P. H., van Duijn, C. M., & Aulchenko, Y. S. (2012). An R package "VariABEL" for genome-wide searching for potentically interacting loci by testing genotypic variance heterogeneity. *BMC genetics*, 13, 4.

Sun, W. (2012) A statistical framework for eQTL mapping using RNA-seq data. *Biometrics,* 68**,** 1-11.

Willer, C.J., Speliotes, E.K., Loos, R.J., Li, S., Lindgren, C.M., Heid, I.M., Berndt, S.I., Elliott, A.L., Jackson, A.U. & Lamina, C. (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature genetics,* 41**,** 25-34.

Yang, J., Loos, R.J.F., Powell, J.E., Medland, S.E., Speliotes, E.K., Chasman, D.I., Rose, L.M., Thorleifsson, G., Steinthorsdottir, V. & Mägi, R. (2012) FTO genotype is associated with phenotypic variability of body mass index. *Nature,* 490**,** 267-272.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., Mcmullen, M.D., Gaut, B.S., Nielsen, D.M. & Holland, J.B. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics,* 38**,** 203-208.

Figure 1. BMI residual box plots and density plots by genotype for rs3120625 and rs7987059. AA, Aa, and aa represent major allele homozygote, heterozygote, and minor allele homozygote, respectively. Residuals were obtained after removing covariate effects and family random effect. rs3120625 is the most significant SNP for mean and variance joint effect. rs7987059 is the most significant SNP for variance only effect.



Table 1. Empirical Type I error/power of association tests for normally distributed traits at significance levels of 0.05 and 0.01. PB: parametric bootstrap.

A. Type I error comparison of family-based tests and tests ignoring family relatedness.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | famLRT | | | | |  | GRAMMAR+ | | |  | LRT | | |
| SNP MAF | level | famLRTMV | famLRTMV(PB) | famLRTM | famLRTV | famLRTV  (PB) |  | LRTMV | LRTM | LRTV |  | LRTMV | LRTM | LRTV |
| 0.44 | 0.05 | 0.054 | 0.049 | 0.052 | 0.053 | 0.046 |  | 0.060 | 0.061 | 0.054 |  | 0.078 | 0.079 | 0.062 |
|  | 0.01 | 0.012 | 0.010 | 0.009 | 0.012 | 0.008 |  | 0.012 | 0.014 | 0.011 |  | 0.023 | 0.022 | 0.018 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.14 | 0.05 | 0.048 | 0.044 | 0.056 | 0.055 | 0.046 |  | 0.057 | 0.056 | 0.061 |  | 0.074 | 0.075 | 0.064 |
|  | 0.01 | 0.012 | 0.009 | 0.012 | 0.011 | 0.008 |  | 0.014 | 0.011 | 0.012 |  | 0.015 | 0.018 | 0.016 |

B. Power comparison of famLRTs and LRTs after GRAMMAR+ transformation in three different scenarios: mean effect, variance effect, mean and variance joint effect.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | famLRT | | | | |  | GRAMMAR+ | | |
| Effect | SNP MAF | level | famLRTMV | famLRTMV(PB) | famLRTM | famLRTV | famLRTV(PB) |  | LRTMV | LRTM | LRTV |
| Mean | 0.44 | 0.05 | 0.573 | 0.521 | 0.692 | 0.048 | 0.042 |  | 0.559 | 0.687 | 0.041 |
|  |  | 0.01 | 0.327 | 0.285 | 0.446 | 0.012 | 0.008 |  | 0.297 | 0.424 | 0.010 |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.14 | 0.05 | 0.266 | 0.236 | 0.334 | 0.053 | 0.050 |  | 0.263 | 0.340 | 0.062 |
|  |  | 0.01 | 0.098 | 0.077 | 0.155 | 0.009 | 0.011 |  | 0.103 | 0.150 | 0.012 |
| Variance | 0.44 | 0.05 | 0.759 | 0.735 | 0.039 | 0.846 | 0.835 |  | 0.711 | 0.062 | 0.816 |
|  |  | 0.01 | 0.555 | 0.508 | 0.007 | 0.663 | 0.634 |  | 0.515 | 0.010 | 0.618 |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.14 | 0.05 | 0.336 | 0.315 | 0.056 | 0.373 | 0.365 |  | 0.298 | 0.055 | 0.358 |
|  |  | 0.01 | 0.147 | 0.144 | 0.012 | 0.188 | 0.175 |  | 0.129 | 0.009 | 0.161 |
| Mean & Variance | 0.44 | 0.05 | 0.919 | 0.915 | 0.580 | 0.837 | 0.819 |  | 0.918 | 0.578 | 0.822 |
|  |  | 0.01 | 0.816 | 0.781 | 0.339 | 0.647 | 0.621 |  | 0.791 | 0.343 | 0.626 |
|  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.14 | 0.05 | 0.533 | 0.512 | 0.282 | 0.376 | 0.362 |  | 0.496 | 0.289 | 0.360 |
|  |  | 0.01 | 0.317 | 0.292 | 0.118 | 0.177 | 0.160 |  | 0.292 | 0.126 | 0.160 |

Table 2. Empirical Type I error/power of association tests for t-distributed and -distributed quantitative traits at significance levels of 0.05 and 0.01. PB: parametric bootstrap.

A. Type I error comparison.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Quantitative Traits | SNP MAF | level | famLRTMV | famLRTMV(PB) | famLRTM | famLRTV | famLRTV(PB) |
| *t*-distributed | 0.44 | 0.05 | 0.083 | 0.045 | 0.055 | 0.090 | 0.044 |
|  |  | 0.01 | 0.022 | 0.008 | 0.012 | 0.031 | 0.010 |
|  |  |  |  |  |  |  |  |
|  | 0.14 | 0.05 | 0.079 | 0.045 | 0.056 | 0.093 | 0.048 |
|  |  | 0.01 | 0.035 | 0.012 | 0.012 | 0.037 | 0.012 |
| -distributed | 0.44 | 0.05 | 0.111 | 0.049 | 0.055 | 0.125 | 0.055 |
|  |  | 0.01 | 0.036 | 0.010 | 0.011 | 0.044 | 0.009 |
|  |  |  |  |  |  |  |  |
|  | 0.14 | 0.05 | 0.104 | 0.056 | 0.056 | 0.105 | 0.053 |
|  |  | 0.01 | 0.030 | 0.011 | 0.013 | 0.040 | 0.013 |

B. Power comparison in three different scenarios: mean effect, variance effect, mean and variance joint effect.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | *t*-distributed | | |  | -distributed | | |
| Effect | SNP MAF | level | famLRTMV  (PB) | famLRTM | famLRTV  (PB) |  | famLRTMV  (PB) | famLRTM | famLRTV  (PB) |
| Mean | 0.44 | 0.05 | 0.460 | 0.633 | 0.055 |  | 0.537 | 0.770 | 0.056 |
|  |  | 0.01 | 0.222 | 0.379 | 0.009 |  | 0.253 | 0.563 | 0.013 |
|  |  |  |  |  |  |  |  |  |  |
|  | 0.14 | 0.05 | 0.190 | 0.311 | 0.048 |  | 0.236 | 0.412 | 0.047 |
|  |  | 0.01 | 0.070 | 0.142 | 0.013 |  | 0.078 | 0.197 | 0.012 |
| Variance | 0.44 | 0.05 | 0.619 | 0.051 | 0.657 |  | 0.449 | 0.043 | 0.584 |
|  |  | 0.01 | 0.408 | 0.010 | 0.433 |  | 0.218 | 0.014 | 0.363 |
|  |  |  |  |  |  |  |  |  |  |
|  | 0.14 | 0.05 | 0.380 | 0.057 | 0.398 |  | 0.173 | 0.046 | 0.256 |
|  |  | 0.01 | 0.184 | 0.013 | 0.202 |  | 0.059 | 0.011 | 0.103 |
| Mean & Variance | 0.44 | 0.05 | 0.799 | 0.513 | 0.675 |  | 0.823 | 0.696 | 0.577 |
|  |  | 0.01 | 0.603 | 0.281 | 0.459 |  | 0.557 | 0.373 | 0.335 |
|  |  |  |  |  |  |  |  |  |  |
|  | 0.14 | 0.05 | 0.504 | 0.275 | 0.384 |  | 0.356 | 0.273 | 0.244 |
|  |  | 0.01 | 0.289 | 0.117 | 0.220 |  | 0.138 | 0.059 | 0.106 |

Table 3. Summary of Framingham Heart Study BMI association analysis results.

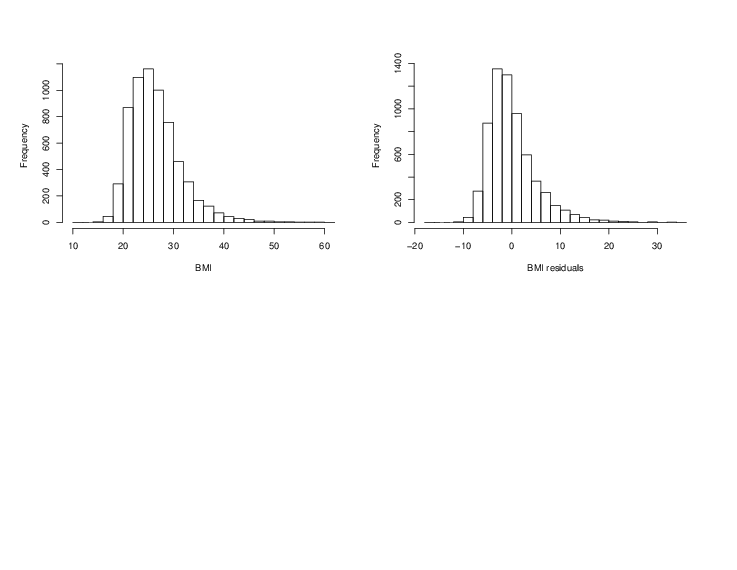
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | p-value using all samples | | |  | p-value using independent samples | | |
| SNP | Chromosome | Position | Nearest Gene | MAF | famLRTMV  (PB) | famLRTM | famLRTV  (PB) |  | LRTMV  (PB) | LRTM | LRTV  (PB) |
| rs3120625 | 1 | 109768889 | *SARS* | 0.344 |  |  |  |  |  |  |  |
| rs7987059 | 13 | 27541625 | *USP12* | 0.181 |  |  |  |  |  |  |  |
| rs9310004 | 3 | 86302812 | *CADM2* | 0.121 |  |  |  |  |  |  |  |
| rs7592497 | 2 | 659958 | *TMEM18* | 0.284 |  |  |  |  |  |  |  |
| rs7621130 | 3 | 86344362 | *CADM2* | 0.120 |  |  |  |  |  |  |  |
| rs352973 | 2 | 142582263 | *LRP1B* | 0.241 |  |  |  |  |  |  |  |
| rs7653885 | 3 | 86344304 | *CADM2* | 0.120 |  |  |  |  |  |  |  |
| rs2889756 | 3 | 186252771 | *CRYGS* | 0.076 |  |  |  |  |  |  |  |
| rs9675886 | 18 | 57969322 | *MC4R* | 0.265 |  |  |  |  |  |  |  |
| rs6781294 | 3 | 86331094 | *CADM2* | 0.120 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |
| rs12328474\* | 2 | 140922100 | *LRP1B* | 0.249 |  |  |  |  |  |  |  |

This table lists the 10 SNPs with smallest p-values of parametric bootstrap based famLRTMV, adjusting for sex, age, and age squared.

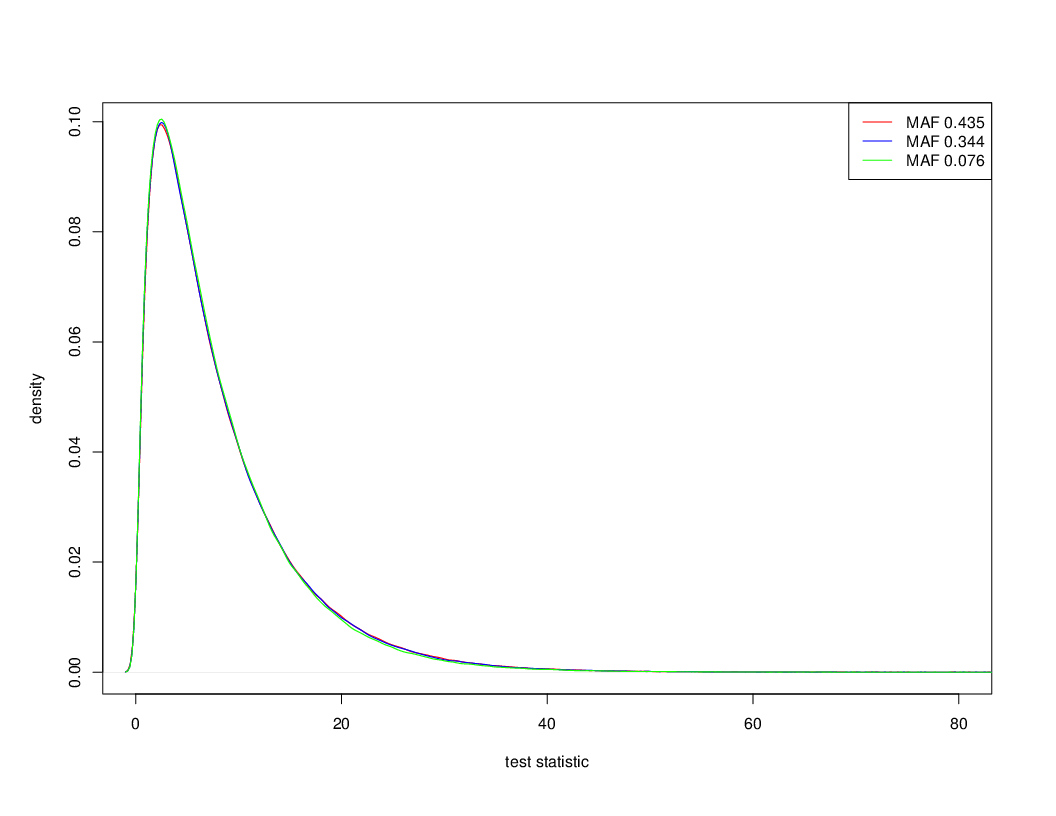
\* denotes the SNP that was identified as a top SNP associated with BMI variability [[Yang et al., 2012](#_ENREF_16)]

## Supplemental Information

Supplemental Figure 1. Histograms of BMI (left) and BMI residuals (right) of all the 6484 individuals. BMI residuals were obtained after removing covariate effects and family random effect.



Supplemental Figure 2. Empirical distribution of parametric bootstrap based famLRTMV test statistic for BMI from the Framingham Heart Study. Empirical distributions of famLRTMV test statistic were constructed using three real tested SNPs with MAF of 0.435, 0.344, and 0.076, respectively. For each SNP, 106 resamplings were performed.



# Chapter 3: Functional Data Analysis Approaches to Mendelian Randomization Study with a Time-varying Exposure

**Title of Journal Article**

Functional Data Analysis Approaches to Mendelian Randomization Study with a Time-varying Exposure

**Name of Journal Proposed for Article Submission**

## Summary

Mendelian randomization (MR) analysis has gained popularity in recent years to make causal inference about an exposure’s effect on a disease outcome from observational epidemiology studies. It is an application of instrumental variable (IV) analysis using genetic variants as IVs. The current practice of MR analysis is to use the exposure’s measure at some arbitrary time point, although many exposures of interest, such as body mass index (BMI), change over one’s lifespan and cannot be adequately captured by a single measure. It remains unknown how to model repeated measures of a time-varying exposure in MR analysis. Here we propose to employ the functional principal component analysis through conditional expectation (PACE) method to recover individual trajectory of the time-varying exposure from longitudinal data, and then to perform IV analysis using the recovered trajectories. We applied the proposed methods to investigate the causal effect of BMI on fasting glucose level using the Framingham Heart Study (FHS) data and observed improved statistical power over standard MR analysis. Simulation studies mimicking the FHS data show that the proposed methods control the Type I error satisfactorily and have higher power than standard methods. We would like to raise the awareness that MR analysis using a single measure of the exposure may lead to statistical power loss and recommend using the proposed methods to take into account time-varying exposures.

KEY WORDS: Mendelian randomization; Instrumental variable; Time-varying exposure; Functional data; FPCA; PACE.

## 1. Introduction

A Mendelian randomization (MR) study stands for an observational epidemiology study that aims to make causal inference about an exposure’s effect on a disease outcome using genetic variants that only affect the disease outcome through the exposure of interest ([Lawlor et al., 2008](#_ENREF_20)). Thanks to the availability of a large number of exposure-associated genetic variants in the form of single nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWAS), MR, a principle originally due to [Katan (1986](#_ENREF_18)), has gained popularity among epidemiologists in recent years to make causal inference of an exposure of interest that cannot be studied through randomized clinical trials (Voight et al, 2012; Holmes et al 2014; Smith et al 2014). For example, Voight et al. (2012) conducted a MR study to investigate the causal effect of plasma high-density lipoprotein (HDL) cholesterol on the risk of myocardial infarction. Holmes et al. (2014) analyzed the causal effects of body mass index (BMI) on cardiometabolic traits and events using MR method.

MR study is an application of instrumental variable (IV) analysis, which is a commonly used method in econometrics to deal with endogeneity. Unobserved confounders can be a cause of endogenetiy ([Wooldridge, 2012](#_ENREF_48)). As depicted by the directed acyclic graph (DAG) in Figure 1A, in a MR study, a genetic variant (G) is the IV that is used to study the effect of the exposure (X) on the outcome variable (Y). The association between X and Y is confounded by some unmeasured confounders (U). To be a valid IV, G must satisfy three assumptions: 1. G is associated with X; 2. G is independent of U that confounds the association between X and Y; 3. G is independent of Y given X and U. The third assumption is also known as the exclusion restriction, meaning that the genetic variant only affects the outcome variable through the exposure. These three assumptions are sufficient for testing the association between the exposure and the outcome variable, while at least another assumption is needed for estimating the causal effect of the exposure on the outcome variable, which is that all the associations in Figure 1A are linear and are not subject to interactions ([Lawlor et al., 2008](#_ENREF_20)).

The most commonly used statistical method for IV analysis with a continuous outcome is two-stage lease squares (2SLS). Specifically, in the first stage, an ordinary least square is performed for the exposure of interest using the IV(s) and measured covariates as the independent variables. In the second stage, another ordinary least square is performed for the outcome variable on the fitted values of the exposure from the first stage and measured covariates. The fitted values of the exposure using IV(s) are independent of unmeasured confounders, thus are independent of the error terms in the second stage linear model. Assuming that the relationship between the exposure and the outcome is linear, the 2SLS estimator is consistent for estimating the causal effect of the exposure on the outcome ([Wooldridge, 2010](#_ENREF_49)). A Wald test based on robust standard error is often used for testing the significance of the causal effect.

Although MR analysis is a very attractive approach to causal inference from observational studies, it is subject to strong assumptions as described above. The first assumption of using genetic variants as IVs is often satisfied by careful selection of SNPs identified by successful GWAS. Due to Mendel’s law of segregation and independence, genotypes are determined by a random process at conception that genetic variants are not associated with unmeasured socioeconomic and behavioral factors that may confound the association between the exposure of interest and outcome variable, satisfying the second assumption. However, there are possible violations of the exclusion restriction assumption of using genetic variants as IVs, including linkage disequilibrium (LD), population stratification, and pleiotropy, i.e., a genetic variant may be associated with other exposures that are associated with the outcome variable ([Lawlor et al., 2008](#_ENREF_20); [VanderWeele et al., 2014](#_ENREF_44)). In addition, weak instrument problem is often encountered in IV analysis, which causes biased causal effect estimates when the exposure-outcome relationship is confounded ([Bound et al., 1995](#_ENREF_4)). In MR studies, a single genetic variant, usually a SNP, only explains a small proportion of the total variation of the exposure, and thus may cause weak IV problem. Genetic score (GS), a weighted count of minor alleles of multiple SNPs based on their estimated effects on the exposure of interest from previous studies, has been used to alleviate the weak IV problem and increase the power in MR studies ([Holmes et al., 2014](#_ENREF_14); [Pierce et al., 2011](#_ENREF_27); Proitsi et al., 2014; [Voight et al., 2012](#_ENREF_45)). Using multiple SNPs has the additional benefit of alleviating the LD, population stratification, and pleiotropy problems that violate the IV analysis assumptions ([Lawlor et al., 2008](#_ENREF_20)).

Time-varying exposure is another challenge in MR studies. Many exposures of interest are time-varying and may have cumulative effect on the outcome variable, for example, HDL cholesterol and BMI. Current MR studies often only use the measured exposure at one time point ([Holmes et al., 2014](#_ENREF_14); [Voight et al., 2012](#_ENREF_45)), for example, the baseline, which, however, may not be representative of the entire trajectory of the time-varying exposure and can underestimate the relationship between the exposure and the outcome (Davis et al., 1990). On the other hand, longitudinal measurements have been collected in many prospective cohort studies, such as the Framingham Heart Study (FHS) ([Splansky et al., 2007](#_ENREF_37)). Figure 1B depicts the scenario that the exposure of interest () is time-varying and longitudinal observations () have been collected at several time points. Taking the FHS as an example, we are interested in investigating whether there is a causal effect of BMI on the fasting glucose level. Figure 2 shows the observed BMI trajectories from the FHS. Different individuals had different numbers of observations scattered at different time points. Cross-sectional data at an arbitrary visit might be used for MR analysis, such as baseline or the last visit. However, the measurement at one time point cannot adequately capture the information of the entire trajectory. We can see from Figure 2 that BMI measures in the FHS varied in one’s lifespan and different individuals had different patterns, for example, almost steady, slowly increasing, slowly decreasing, or going up and down. Although it sounds appealing to incorporate longitudinal measurements of the exposure in MR analysis to capture the time-varying information, it remains unknown how to do so, especially when the longitudinal measurements are irregular.

In this work, we propose to employ a functional data analysis based framework for IV analysis of a time-varying exposure in MR studies. More specifically, we propose to recover the trajectory of the time-varying exposure for each subject using the functional principal component analysis through conditional expectation (PACE) method in the first step, which was developed for modeling irregular and sparse longitudinal data (Yao et al. 2005). The second step is to perform IV analysis using the recovered trajectories from the first step. We propose two methods for IV analysis. The first method is to summarize each trajectory by integration and then perform standard IV analysis for the summarized trajectories using the 2SLS. The second method is to use two-stage functional linear regression (2SFLR) to capture the time-varying effect of genetic variants on the exposure of interest in IV analysis. The details of the proposed methods are described in Section 2. In Section 3, we applied the proposed methods to the FHS data to analyze the effect of BMI on fasting glucose level. We conducted simulation studies mimicking the FHS real data to evaluate the properties of the proposed methods in Section 4. Finally, we conclude with a brief discussion in Section 5. To the best of our knowledge, this is the first work on incorporating time-varying information of an exposure in MR study.

## 2. Methods

*2.1 Notations*

We consider a cohort study with a total of subjects. Let be the outcome variable of subject at time and be the time-varying exposure of interest of subject at time . Let denote the genetic score, i.e., the IV, of subject and be the covariate vector of subject . In this study, we only consider the case that both and are continuous variables.

The ideal data would satisfy that and for , meaning all the subjects have the same number of observations at the same time points. However, we almost never obtain perfect balanced longitudinal measurements from observational studies involving human subjects in reality. For example, in our motivating example of the FHS data, some individuals had data collected in all the seven clinical visits, while others had data collected in six or five clinical visits, or even fewer. In the same clinical visit, different individuals were at different ages (Figure 2).

*2.2 Standard IV analysis for a time-varying exposure*

To study a time-varying exposure for which sparse longitudinal measurements have been collected, [Sánchez et al. (2010](#_ENREF_32)) suggested grouping into time intervals with certain length, for example, five-year interval. If a subject has multiple measurements in the same time interval, then the average value can be used. Let be the time-varying exposure at the th time interval, which may be an observed value or the mean of multiple observed values in the same time interval. Then a 2SLS analysis can be conducted in each time interval by fitting a linear model for in the first stage: to obtain the fitted value , where is the error term. The second stage linear model is: , where is the error term. To test the null hypothesis that the time-varying exposure has no effect on the outcome variable (), we can test using a Wald test with robust standard error in each 2SLS analysis and obtain the p-value . Then the minimum p-value, ), can be compared with the Bonferroni-corrected significance threshold . One of the limitations of the minP method is that the sample size in each interval may vary a lot when the outcome variable is measured at different time points; see Web Table 1 for an illustration of the FHS data, to be detailed in Section 3.

*2.3 New Methods: IV analysis for a time-varying exposure using functional data analysis approaches*

Functional data analysis has been developed as a powerful and flexible tool for modeling sparse longitudinal data with irregular measurements ([Müller, 2009](#_ENREF_24); [Yao et al., 2005](#_ENREF_52)). Time-varying exposures, for example, BMI, are intrinsically continuous, even though only a limited number of measurements are taken at certain time points, leading to sparse longitudinal observations. Therefore, we propose to employ functional data analysis tools to perform IV analysis for time-varying exposures.

The first step is to recover the individual trajectory of the time-varying exposure using the PACE method developed by Yao et al. (2005). Briefly, the sparse longitudinal observations of each subject are modeled as noisy sampled points from an underlying trajectory. The collection of trajectories of all the subjects are assumed to be independent and have a mean function and a covariance function in a closed time interval . The covariance function can be expanded using the eigendecomposition as , where ’s are eigen-values () and ’s are eigen-functions. By the Karhunen-Loève theorem, a random curve can be expressed as , where is the th functional principal component (FPC) score of subject (Yao et al., 2005). has mean of 0 and variance of . The model was further extended by Yao et al. (2005) to incorporate additive measurement errors as , where the measurement error follows the classical measurement error assumption of and ([Carroll et al., 2006](#_ENREF_8)). To obtain the estimated functions and , all the observations are pooled together. Local linear smoother is used to obtain the estimated mean function . The sample covariance function is smoothed using a two-dimensional local smoother to obtain . When measurement errors is taken into account, the sample covariance function is smoothed using a local linear smoother in the direction of the diagonal and a local quadratic smoother in the direction orthogonal to the diagonal. Then eigendecomposition is conducted for after discretization. The traditional FPC score calculation by numerical integration does not work well with sparse data contaminated with measurement errors. By assuming that and are joint normal, Yao et al. (2005) suggested that the best prediction for is the conditional expectation and can be estimated by plugging in parameter estimates obtained in the previous steps. Finally, each trajectory can be recovered by using the leading *K* eigen-functions as for . The number of eigen-functions can be chosen by using either the fraction of variance explained (FVE) or AIC-type criteria (Yao et al., 2005). The PACE method for sparse longitudinal data has been implemented in R package “PACE”.

The next step is to perform IV analysis for the time-varying exposure using the PACE-recovered trajectories. We assume that the time-varying exposure has accumulative effect on the outcome variable. Two methods are proposed for conducting IV analysis of the recovered trajectories, as illustrated in Web Figure 1. The first method is to summarize each trajectory by integration and then perform standard IV analysis for the integrated trajectories, denoted as PACE+2SLS. Specifically, let , where is the lower bound of the time interval , is the time that the outcome variable of subject is observed. is the cumulative value of the time-varying exposure, mimicking dosage. Then the effect of the cumulative exposure on the outcome variable can be assessed using 2SLS. In detail, a linear model is fitted in the first stage: to obtain the fitted value , where is the error term. Then the accumulative effect is tested and estimated in the second stage linear regression model , where is the error term and we test the null hypothesis using a Wald test with robust standard error.

On the other hand, gene expression is time-varying in the development process. Therefore, we would expect that the effect of genetic variants on the exposure is also time-varying. To model this phenomenon, we propose the second method to conduct IV analysis using a two-stage functional linear regression, denoted as PACE+2SFLR. To avoid notation confusion, let . In the first stage of the 2SFLR, a functional linear model is fitted for the time-varying exposure: , , to obtain the fitted . For computational purpose, the time varying functions need to be expanded using standard basis systems, for example, the B-spline basis system ([Ramsay and Silverman, 2005](#_ENREF_29)). Functional linear models, including the above function-on-scalar model, have been implemented in the R package “fda”. Then the effect of GS-instrumented exposure on the outcome is assessed using the second stage linear model: . Let be the fitted residuals. To test the null hypothesis that the time-varying exposure has no effect on the outcome variable () and to accommodate the adjustment for covariates, we propose to use a parametric bootstrap based -test. Specifically, a reduced linear model is fitted and let be the fitted residuals of the reduced model. The -test statistic is calculated as . Parametric bootstrap can be used to construct the distribution of the -test statistic under the null hypothesis as follows: 1. obtain the fitted outcomes and the fitted residuals from the reduced model ; 2. permute ’s to generate ’s and construct for ; 3. fit the second-stage linear model and the reduced linear model to calculate by replacing with for ; 4. repeat steps 2 and 3 *B* times and the p-value of the -test is given by .

With a time-varying exposure of interest, the effect estimates using different IV analysis methods have different interpretations. The focus here is to test the causal effect of a time-varying exposure on an outcome variable, instead of estimation. This is in line with that the main focus of MR studies is to identify causal risk factors, not to obtain precise effect estimates ([Burgess, 2013](#_ENREF_5)). In addition, stronger assumptions are needed for effect estimation in IV analysis ([Lawlor et al., 2008](#_ENREF_20)), which are rarely satisfied in real data analysis.

## 3. Application to the FHS Data

The FHS is a prospective cohort study conducted since 1948 to identify risk factors for cardiovascular disease and three generations have been recruited: the Original Cohort, the Offspring Cohort, and the Third Generation Cohort ([Splansky et al., 2007](#_ENREF_37)). The genotype and phenotype data of the FHS were downloaded from NCBI dbGaP (http://www.ncbi.nlm.nih.gov/gap). The FHS is one of the studies included in the MR analysis for causal effect of BMI on cardiometabolic traits and events (Holmes et al., 2014). However, only baseline BMI information was used even though longitudinal measurements have been collected. We applied the proposed methods to perform MR analysis of the causal effect of longitudinal BMI on fasting glucose level and compared the results with those of standard MR analysis with the BMI at a single time point. We extracted the longitudinal BMI data of unrelated individuals in the FHS Offspring Cohort, which is the largest cohort among the three generations with both genotype information and longitudinal phenotype information collected from seven clinical exams ([Cupples et al., 2009](#_ENREF_11)). We selected the age interval from 25 to 75 years old so that at least 50 BMI observations were collected at each age. We used the last fasting glucose level collected before taking any diabetes related medications in the 25-75 years old interval as the outcome variable. Individuals with at least two BMI measurements before the outcome variable were included in the analysis, resulting in a total sample size of 1709. Following Holmes et al. (2014), we used the BMI GS constructed using 14 SNPs as the IV. The 14 SNPs were selected carefully and checked for possible violations of the MR study assumptions, with detailed information in Table S2 of Holmes et al. (2014). We extracted the 14 SNPs from the FHS Candidate Gene Association Resource (CARe) study. We included sex and age, at which the outcome variable information was collected, as covariates in the analysis. The *F*-test statistic for the association between the GS and longitudinal BMI using a linear mixed model (LMM) was 13.98, higher than the suggested weak IV threshold of 11 ([Pierce et al., 2011](#_ENREF_27)). In addition, association analysis between the GS and the last fasting glucose level adjusting for accumulative BMI had a p-value of 0.335, suggesting no violation of the exclusion restriction assumption of IV analysis.

The MR analysis results are shown in Table 1. The minP method using 5-year intervals had a Bonferroni-corrected p-value of 0.322. In comparison, the PACE+2SLS method had a p-value of 0.0872 and the PACE+2SFLR method had a p-value of 0.266. When the minP method was used, sample sizes, IV strength (reflected by the *F*-test statistics of the effect of GS on BMI), and p-values varied dramatically in the ten age intervals (Web Table 1). In contrast, the PACE procedure summarized irregular longitudinal observations into one piece of information for subsequent IV analysis. Although the individual BMI profiles showed substantial variations (Figure 2), the selected top three FPCs estimated by the PACE method captured 86.2%, 12.2%, and 1.4% of the total variation, respectively, with a total of 99.8% (Web Figure 4). Wed Figure 3 shows the top three eigenfunctions. The first eigenfunction represents an almost constant upward shift of BMI from the mean curve. The second eigenfunction shows the pattern of lower BMI values than the mean curve prior to 50 years old and higher BMI values than the mean curve after 50 years old. The third eigenfunction is a concave curve representing a BMI increase from 25 to 50 years old followed by a decrease from 50 to 75 years old. The individual plots of observed BMI versus PACE-predicted trajectories show that the PACE method recovered the individual trajectories very well (Web Figure 2). The recovered trajectory was in good concordance with the observed values for the two randomly selected individuals and the individual with the largest FPC2 score. For the individual with the largest FPC3 score, the PACE-recovered trajectory shrank the extreme BMI values toward the mean curve, while maintaining its unique trend. All of these confirm that the PACE method is versatile and powerful in synchronizing irregular longitudinal observations into a smooth trajectory for each individual. In the context of gene-environment interactions, Wei et al (2014) combined the PACE method with the functional logistic regression model to detect gene by longitudinal environmental exposure interaction influencing the risk of complex disease.

Of note, a previous MR analysis using a single measure of BMI identified a significant causal effect of BMI on fasting glucose level with a total sample size of 34,538 (Holmes et al., 2014). Direct analysis of the association between the longitudinal BMI and fasting glucose level using linear regressions showed very significant results (Table 1). However, none of the IV analyses we performed using longitudinal BMI data showed a significant result. The analysis was possibly underpowered due to two reasons: the sample size was small and the GS was not a strong IV. Simulation studies in the next section confirmed the lack of power in our real data analysis. In addition, we also conducted MR analysis of the effect of BMI on fasting glucose level using cross-sectional data for comparison. We used standard 2SLS to analyze BMI and fasting glucose level data collected concurrently from each of the clinical visits. The results are shown in Web Table 2. Although the sample sizes were comparable across visits, the IV analysis results varied dramatically. The Bonferroni-corrected minimum p-value of the cross-sectional analysis was 0.195. This suggests that the current practice of MR analysis using a single measure of the exposure at some arbitrary time point may miss the critical time windows when the exposure has a causal effect on the disease outcome, leading to loss of statistical power.

## 4. Simulation Study

To evaluate Type I error rate and statistical power of the methods we proposed for conducting IV analysis with a time-varying exposure, we performed simulation studies mimicking the FHS data. We simulated the time-varying exposure, outcome variable, IV, and covariates using parameter estimates from the FHS data analysis. Specifically, the time-varying exposure of subject at time was simulated using the model: , where and . denotes the age when the outcome variable was observed, randomly sampled from . The genetic score and a hypothesized unmeasured confounder were simulated from normal distributions, and , respectively. Sex was simulated from Bernoulli(0.5). The residual vector was resampled from the residual vectors estimated from the first stage of 2SFLR in the FHS data analysis. We let , the estimated mean function of FHS BMI from the PACE procedure. and are equal to the estimated time-varying coefficients of GS and sex from the first stage of 2SFLR in the FHS data, respectively. We also let the coefficient of the unmeasured confounder be time-varying: . The outcome variable was simulated as: . The estimate of from the real data was 0.053. The residual was simulated from . To simulate irregular measured exposures, we randomly sampled two to five longitudinal observations of the time-varying exposure at randomly selected time points for each subject as the observed data. For Type I error evaluation, we let and considered different strength of the IV: , , and We also included the linear regression based association analysis of the effect of on for comparison. To investigate the statistical power of different methods, we conducted simulation studies using different strength of the IV as in the Type I error evaluation, and also different causal effect sizes of on , including , and . We used the sample size of 1000 for simulation studies. 5000 replication were conducted for Type I error evaluation and 1000 replications were conducted for power evaluation. 5-year intervals were used for the minP method. We selected the number of FPCs in the PACE procedure based on the FVE of at least 95%.

Table 2 shows the empirical Type I error rates from simulation studies. When there were unmeasured confounders, linear regression based association analysis had substantial Type I error inflation. In contrast, IV analysis methods were able to control the Type I error rate except the uncorrected minP method, confirming that Bonferroni correction is necessary when using the minP method to test the causal effect of a time-varying exposure on an outcome variable. When the association between GS and the time-varying exposure was weak, reflected by the *F*-test statistic on the effect of GS on longitudinal observations using a LMM, the empirical Type I error rates of all three methods, Bonferroni corrected minP, PACE+2SLS, and PACE+2SFLR, were much lower than the nominal level. As the strength of the IV increased, the empirical Type I error rates of all three methods gradually approached the nominal level, with the PACE+2SFLR method being the most conservative.

Table 3 shows the empirical powers of the methods that were able to control the Type I error. We can see that the power of all three methods increased when either the IV strength or the effect of on increased. The power of the PACE+2SLS method was always higher than the Bonferroni-corrected minP method in all the simulation scenarios. The PACE+2SFLR method had the lowest power among the three methods when the simulated effect of GS on or . However, when the simulated effect of GS on increased to , the power of PACE+2SFLR method surpassed the Bonferroni-corrected minP method. When the effect size further increased to , the PACE+2SFLR method had the highest power.

In addition to effect size, we further evaluated the impact of sample size using simulation studies. Following the same simulation setup described above, we investigated the empirical Type I error and power of sample size 2000 and 5000 when the simulated effect of GS on was fixed at and the simulated effect of on was . Type I error results are shown in Web Table 3 and power results are shown in Web Table 4. We see substantial power increase of the Bonferroni-corrected minP method and the PACE+2SLS method when the sample size increased. However, the sample size did not have much effect on the power of the PACE+2SFLR method when the effect size remained small. We also performed simulation study to investigate if the Type I error and power were affected when the longitudinal observations were contaminated with measurement error: we let the observed value be , where is the measurement error, simulated from with 1.17 being the estimated standard deviation of measurement error from the PACE procedure in the FHS data analysis. We did not observe any effect introduced by the measurement error (results not shown), likely because the PACE procedure in the first stage effectively alleviated the impact of the measurement errors.

## 5. Discussion

We have proposed two functional data analysis based methods for conducting IV analysis when longitudinal observations have been collected for a time-varying exposure and the outcome variable is continuous. We applied the proposed methods to the FHS data to investigate the causal effect of BMI on fasting glucose level using a GS as the IV. Simulation studies showed that the PACE+2SLS method always outperformed the standard Bonferroni-corrected minP method, which groups longitudinal observations into pre-determined age intervals. The other functional data analysis based method, PACE+2SFLR, showed advantage when a very strong IV was used.

IV analysis techniques have been widely used in the econometrics literature for causal inference from observational studies. In recent years, epidemiologists started to apply IV analysis in observational epidemiology studies to draw causal inference by using genetic variants as IVs, which is known as MR analysis. Although strong assumptions are needed for a valid MR analysis ([Lawlor et al., 2008](#_ENREF_20)), it is a useful technique for reducing bias and narrowing down risk factors from observational studies for further investigation. Many exposures are time-varying and longitudinal observations have been collected from observational epidemiology studies. However, current MR studies have not made use of the longitudinal data. To incorporate the information from longitudinal data, we developed two IV analysis approaches using functional data analysis techniques. To our knowledge, we are the first to address the challenge of using longitudinal observations in MR studies.

As demonstrated in the FHS BMI data analysis, the PACE method is flexible and powerful for analyzing sparse longitudinal data. It standardizes irregular longitudinal data into functional data for subsequent IV analysis. Although 2SFLR is able to capture the time-varying effect of IV on the exposure variable, it is a complicated method involving estimation of more parameters, resulting in efficiency loss. Therefore, a very strong IV is needed for 2SFLR to reach high power, which makes the method less practical in MR analysis because at present it is rare to find genetic variants as such strong IVs that can get high power using the 2SFLR. In contrast, 2SLS using summary statistic of functional data is a simpler method, gains efficiency and shows higher power than 2SFLR in most simulation scenarios. Functional data analysis based methods are preferred over the Bonferroni-corrected minP method because at least one of the two functional data analysis based methods has higher power than the minP method in all the scenarios. We recommend using PACE+2SLS to conduct MR analysis for a time-varying exposure with longitudinal data unless a very strong IV is identified so that PACE+2SFLR can also be employed.

We did not identify a significant causal effect of longitudinal BMI on fasting glucose level using the FHS data. The analysis was underpowered due to both small sample size and small effect size of the IV. Based on the simulated data resembling the effect size of the FHS data with a sample size of 2000, the highest empirical power was only 13.6% by using the PACE+2SLS method (Web Table 4). This is in good agreement with that the smallest p-value in the FHS data analysis was obtained by the PACE+2SLS method, which was 0.0872. Therefore, the FHS data analysis result does not indicate a controversial result from the previous MR study in which a significant causal effect of BMI on fasting glucose level was identified (Holmes et al., 2014). On the other hand, the standard MR analysis of the FHS data resulted in even less significant p-values (Web Table 2), suggesting that Holmes et al. were able to compensate the power loss due to using a single measure of BMI by substantially increasing the sample size (1709 in our analysis versus 34538 therein).

In this paper, we limited the IV analysis for a time-varying exposure to the case that the outcome variable is continuous. Disease outcomes in epidemiology studies are often binary. IV analysis with a binary outcome is more challenging because of the nonlinear relationship between the exposure and the outcome variable (Dai et al., 2014). Future work on developing methods for investigating the causal effect of a time-varying exposure on a binary outcome will provide a more complete framework for conducting IV analysis for time-varying exposures.

ACKNOWLEDGEMENTs

This research was supported by the National Institutes of Health grants R01CA169122, R01HL116720 and R21HL126032. The authors declare no conflict of interest. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL-64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University.

## References

Bound, J., Jaeger, D.A., and Baker, R.M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443-450.

Burgess, S. (2013). Identifying the odds ratio estimated by a two‐stage instrumental variable analysis with a logistic regression model. *Statistics in Medicine* 32, 4726-4747.

Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models: a Modern Perspective.* New York: Chapman & Hall.

Cupples, L.A., Heard-Costa, N., Lee, M., and Atwood, L.D. (2009). Genetics analysis workshop 16 problem 2: the Framingham Heart Study data. *BMC Proceedings.* (Vol. 3, No. Suppl 7, p. S3) BioMed Central Ltd.

Davis, C. E., Rifkind, B. M., Brenner, H., & Gordon, D. J. (1990). A single cholesterol measurement underestimates the risk of coronary heart disease: an empirical example from the Lipid Research Clinics Mortality Follow-up Study. *JAMA* 264, 3044-3046.

Holmes, M.V., Lange, L.A., Palmer, T., Lanktree, M.B., North, K.E., Almoguera, B., Buxbaum, S., Chandrupatla, H.R., Elbers, C.C., and Guo, Y. (2014). Causal effects of body mass index on cardiometabolic traits and events: a mendelian randomization analysis. *The American Journal of Human Genetics* 94, 198-208.

James Y. Dai, Kwun Chuen Gary Chan and Li Hsu (2014) Testing concordance of instrumental variable effects in generalized linear models with application to Mendelian randomization. *Statistics in Medicine* 33, 3986–4007.

Katan, M. (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. *The Lancet* 327, 507-508.

Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics In Medicine* 27, 1133-1163.

Müller, H.-G. (2009). Functional modeling of longitudinal data. In: *Longitudinal Data Analysis (Handbooks of Modern Statistical Methods)* New York: Wiley.

Pierce, B.L., Ahsan, H., and VanderWeele, T.J. (2011). Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology* 40, 740-752.

Proitsi, P., Lupton, M. K., Velayudhan, L., Newhouse, S., Fogh, I., Tsolaki, M., ... & GERAD1 Consortium. (2014). Genetic Predisposition to Increased Blood Cholesterol and Triglyceride Lipid Levels and Risk of Alzheimer Disease: A Mendelian Randomization Analysis. *PLoS Medicine* 11, e1001713.

Ramsay, J.O., and Silverman, B.W. (2005). *Functional Data Analysis.* New York: Springer-Verlag.

Sánchez, B.N., Hu, H., Litman, H.J., and Téllez-Rojo, M.M. (2010). Statistical methods to study timing of vulnerability with sparsely sampled data on environmental toxicants. *Environmental Health Perspectives* 119, 409-415.

Smith JG, Luk K, et al (2014) Association of Low-Density Lipoprotein Cholesterol–Related Genetic Variants With Aortic Valve Calcium and Incident Aortic Stenosis. *JAMA* 312:1764-1771.

Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B., Fox, C.S., Larson, M.G., and Murabito, J.M. (2007). The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *American Journal of Epidemiology* 165, 1328-1335.

VanderWeele, T.J., Tchetgen, E.J.T., Cornelis, M., and Kraft, P. (2014). Methodological Challenges in Mendelian Randomization. *Epidemiology* 25, 427-435.

Voight, B.F., Peloso, G.M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M.K., Hindy, G., Hólm, H., Ding, E.L., and Johnson, T. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet* 380, 572-580.

Wei P, Tang H, Li D (2014) Functional logistic regression approach to detecting gene by longitudinal environmental exposure interaction in a case-control study. *Genetic Epidemiology* 38, 638-651.

Wooldridge, J.M. (2012). Introductory Econometrics: A Modern Approach. Cengage Learning.

Wooldridge, J.M. (2010). Econometric Analysis of Cross Section and Panel Data. MIT press.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 100, 577-590.

Table 1. P-values of testing the effect of longitudinal BMI on fasting glucose level using the FHS data. 5-year intervals are used for the Bonferroni-corrected minP method.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| IV analysis | | |  | Linear regression | |
| minP (Bonferroni) | PACE+2SLS | PACE+2SFLR |  | minP (Bonferroni) | PACE+LR |
| 0.322 | 0.0872 | 0.266 |  | < 2E-16 | < 2E-16 |

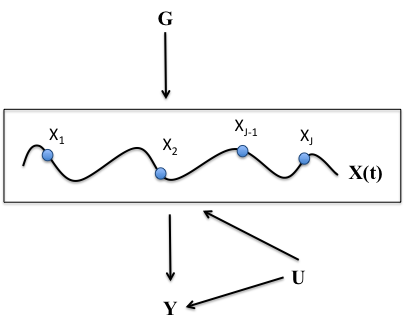
Table 2. Empirical Type I error rates of different methods at the significance level of 0.05.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Effect of GS on | mean *F-*test statistic  from LMM | IV analysis | | | |  | Linear regression | | |
| corrected  minP | uncorrected  minP | PACE  +2SLS | PACE  +2SFLR |  | corrected  minP | uncorrected  minP | PACE  +LR |
|  | 8.56 | 0.0004 | 0.0248 | 0.0146 | 0.0004 |  | 0.2888 | 0.6050 | 0.4202 |
| 2 | 31.1 | 0.0054 | 0.1076 | 0.0368 | 0.0040 |  | 0.2790 | 0.5966 | 0.4112 |
| 4 | 121 | 0.0280 | 0.1928 | 0.0480 | 0.0214 |  | 0.2594 | 0.5726 | 0.3874 |
| 6 | 271 | 0.0392 | 0.2174 | 0.0500 | 0.0304 |  | 0.2284 | 0.5446 | 0.3500 |

Table 3. Empirical power of different methods at the significance level of 0.05.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Effect of GS on | mean *F-*test statistic  from LMM |  | Effect of on |  | IV analysis | | |
|  |  | corrected  minP | PACE  +2SLS | PACE  +2SFLR |
|  | 8.65 |  |  |  | 0.006 | 0.031 | 0.001 |
|  |  |  | 0.024 | 0.073 | 0.007 |
|  |  |  | 0.141 | 0.246 | 0.021 |
|  |  |  |  |  |  |  |  |
| 2 | 31.1 |  |  |  | 0.048 | 0.078 | 0.030 |
|  |  |  | 0.133 | 0.239 | 0.064 |
|  |  |  | 0.538 | 0.690 | 0.276 |
|  |  |  |  |  |  |  |  |
| 4 | 121 |  |  |  | 0.150 | 0.232 | 0.188 |
|  |  |  | 0.572 | 0.715 | 0.664 |
|  |  |  | 0.988 | 0.997 | 0.994 |
|  |  |  |  |  |  |  |  |
| 6 | 270 |  |  |  | 0.348 | 0.497 | 0.539 |
|  |  |  | 0.883 | 0.966 | 0.967 |
|  |  |  | 1 | 1 | 1 |

Figure 1. DAG for a MR study. G is a genetic variant (an instrumental variable (IV)), X is the exposure of interest, Y is the outcome variable, and U represents unmeasured confounders. (A). The exposure of interest X is constant or measured only at one time point. (B). The exposure of interest X(t) is time-varying. represent observed values of the time-varying exposure at J time points.



**B.**

**A.**

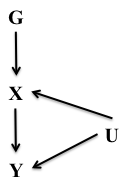
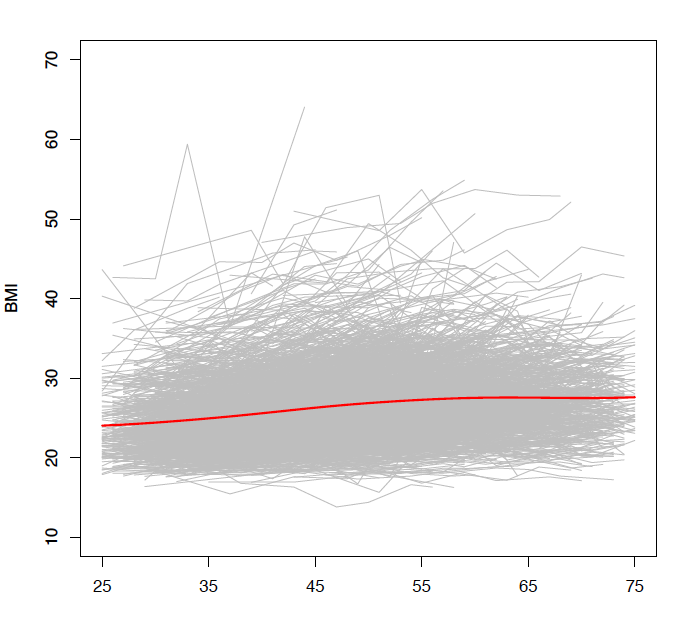


Figure 2. Longitudinal BMI observations from the FHS. Each gray line corresponds to the BMI observations of an individual included in the MR analysis. The red line shows the estimated mean function of BMI from the PACE procedure.



## Supplementary Materials

Web Table 1. The effect of longitudinal BMI in each age interval on the last fasting glucose level using the minP method. \*: Bonferroni-corrected minimum p-value.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age interval |  | sample size |  | IV analysis | |  | Linear regression  P-value |
|  |  | P-value | *F*-test statistics of the effect of GS on BMI |  |
| 25-29 |  | 422 |  | 0.465 | 2.24 |  | 2.50E-06 |
| 30-34 |  | 631 |  | 0.032 | 14.37 |  | 1.69E-15 |
| 35-39 |  | 901 |  | 0.074 | 15.10 |  | 2.44E-20 |
| 40-44 |  | 1169 |  | 0.089 | 4.21 |  | 5.18E-33 |
| 45-49 |  | 1344 |  | 0.110 | 9.42 |  | 9.16E-26 |
| 50-54 |  | 1362 |  | 0.128 | 5.81 |  | 4.30E-22 |
| 55-59 |  | 1160 |  | 0.245 | 4.27 |  | 1.21E-19 |
| 60-64 |  | 834 |  | 0.287 | 8.18 |  | 4.94E-15 |
| 64-69 |  | 538 |  | 0.833 | 1.94 |  | 5.29E-11 |
| 70-75 |  | 139 |  | 0.561 | 0.33 |  | 2.32E-04 |
| - |  | - |  | 0.322\* | - |  | < 2E-16\* |

Web Table 2. The effect of BMI on fasting glucose level using cross-sectional data. Age and sex are the covariates adjusted in the analysis. Fasting glucose level data are not available for clinical visit 1 and 2. Analysis is only conducted for clinical visit 3 to 7. \*: Bonferroni-corrected minimum p-value.

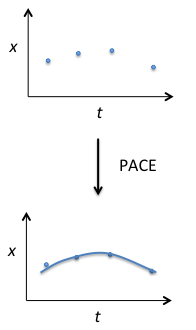
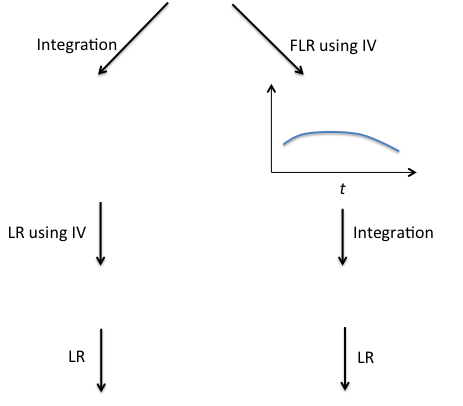
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| visit |  | sample size |  | age median |  | IV analysis | |  | Linear regression  P-value |
|  |  |  | P-value | *F*-test statistics of the effect of GS on BMI |  |
| 3 |  | 1423 |  | 47 |  | 0.353 | 9.08 |  | < 2E-16 |
| 4 |  | 1498 |  | 49 |  | 0.601 | 12.07 |  | < 2E-16 |
| 5 |  | 1574 |  | 53 |  | 0.408 | 11.83 |  | < 2E-16 |
| 6 |  | 1528 |  | 57 |  | 0.039 | 6.82 |  | < 2E-16 |
| 7 |  | 1539 |  | 60 |  | 0.136 | 11.13 |  | < 2E-16 |
| - |  | - |  | - |  | 0.195\* | - |  | < 2E-16\* |

Web Table 3. Empirical Type I error rate using different sample sizes at the significance level of 0.05. The simulated effect of GS on is

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| sample size | mean *F-*test statistic  from LMM | IV analysis | | | |
| corrected  minP | uncorrected  minP | PACE  +2SLS | PACE  +2SFLR |
| 1000 | 8.56 | 0.0004 | 0.0248 | 0.0146 | 0.0004 |
| 2000 | 15.4 | 0.0008 | 0.0318 | 0.0180 | 0.0006 |
| 5000 | 37.8 | 0.0092 | 0.0964 | 0.0440 | 0.0012 |

Web Table 4. Empirical power using different sample sizes at the significance level of 0.05. The simulated effect of GS on and the simulated effect of on is .

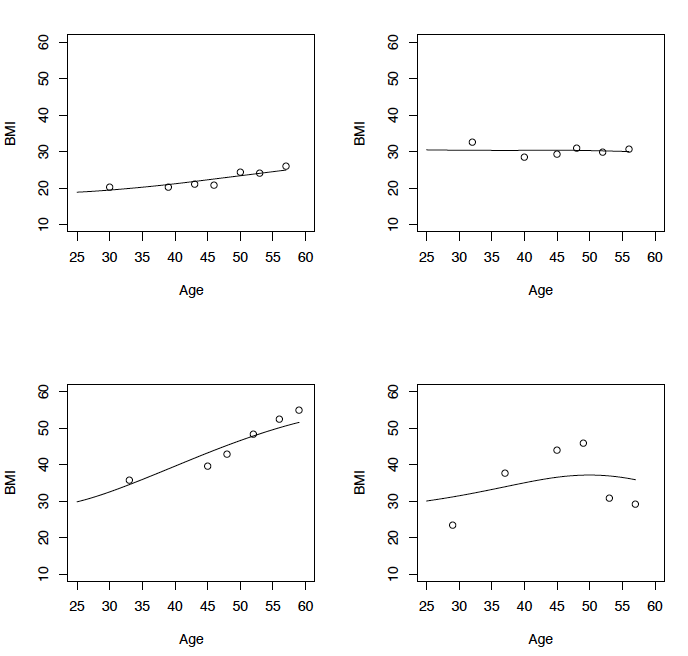
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sample size | mean *F-*test statistic  from LMM | IV analysis | | |
| corrected  minP | PACE  +2SLS | PACE  +2SFLR |
| 1000 | 8.65 | 0.024 | 0.073 | 0.007 |
| 2000 | 15.4 | 0.038 | 0.136 | 0.001 |
| 5000 | 37.8 | 0.166 | 0.309 | 0.004 |

Web Figure 1. Illustration of the two functional data analysis based approaches for IV analysis of a time-varying exposure. The diagram on the left illustrates the PACE+2SLS method, while the one on the right illustrates the PACE+2SFLR method.

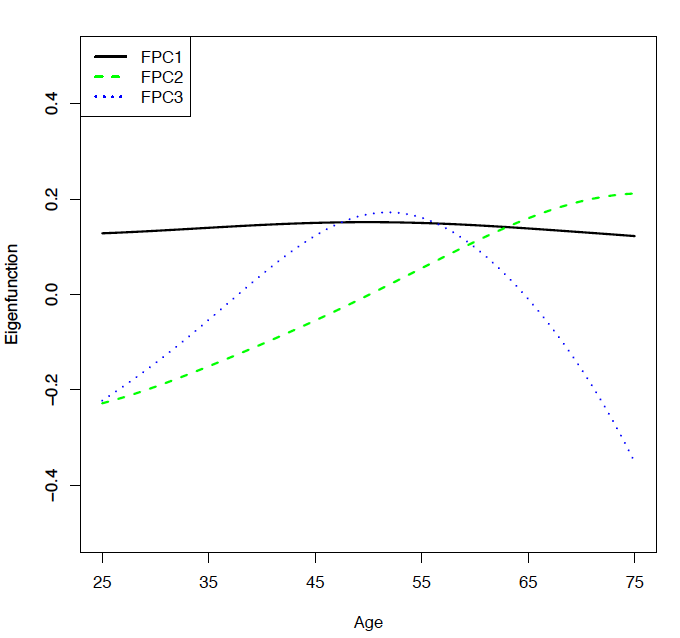
Effect of on

Effect of on

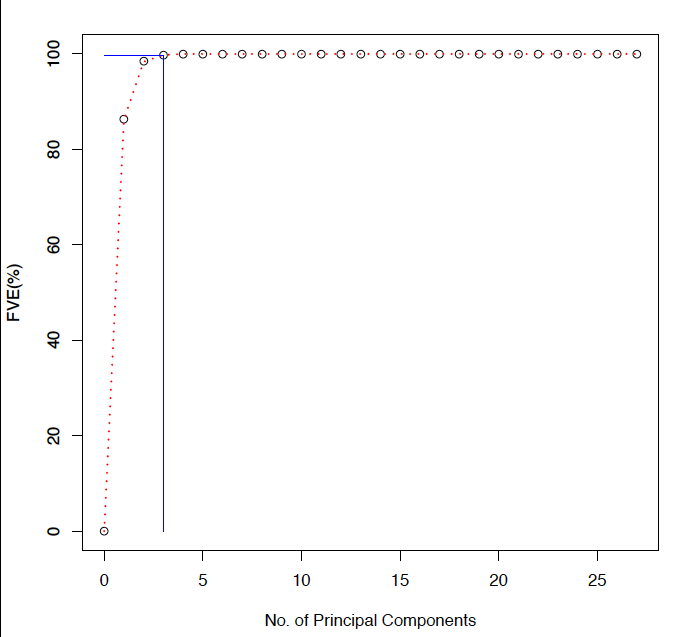
Web Figure 2. Observed BMI values vs. PACE-predicted trajectories, represented by circles and solid lines, respectively. The two plots on the top display the BMI profiles of two randomly selected individuals. The lower-left plot displays the BMI profile of the individual with largest FPC2 score. The lower-right plot displays the BMI profile of the individual with largest FPC3 score.



Web Figure 3. The three leading functional principal component (FPC) functions of the FHS longitudinal BMI observations estimated using the PACE method.



Web Figure 4. Fraction of variance explained (FVE) by the number of functional principal components (FPCs). The top three FPCs explain 99.8% of the total variation.



# Chapter 4: Mendelian Randomization Analysis of a Time-varying Exposure for Binary Disease Outcomes using Functional Data Analysis Methods

**Title of Journal Article**

Mendelian Randomization Analysis of a Time-varying Exposure for Binary Disease Outcomes using Functional Data Analysis Methods

**Name of Journal Proposed for Article Submission**

## Abstract

A Mendelian randomization (MR) analysis is to analyze the causal effect of an exposure variable on a disease outcome from observational studies by using genetic variants that affect the disease outcome only through the exposure variable. This method has recently gained popularity among epidemiologists given the success of genetic association studies. Many exposure variables of interest are time-varying, for example, body mass index (BMI). Although longitudinal data have been collected in many cohort studies, current MR studies only use one measurement of a time-varying exposure variable, which cannot adequately capture the long-term time-varying information. We propose to use the functional principal component analysis method to recover the underlying individual trajectory of the time-varying exposure from the sparsely and irregularly observed longitudinal data, and then conduct MR analysis using the recovered curves. We further propose two MR analysis methods are proposed. The first assumes a cumulative effect of the time-varying exposure variable on the disease risk, while the second assumes a time-varying genetic effect and employs functional regression models. We focus on statistical testing for a causal effect. Our simulation studies mimicking the real data show that the proposed functional data analysis-based methods incorporating longitudinal data have substantial power gain as compared with standard MR analysis using only one measurement. We used the Framingham Heart Study (FHS) data to demonstrate the promising performance of the new methods as well as inconsistent results by the standard MR analysis that relies on a single measurement of the exposure at some arbitrary time point.

## Introduction

Complex diseases are influenced by multiple genetic factors, as well as behavioral and environmental factors, and their interactions. In the past ten years, tremendous progress has been made in genetic studies for complex diseases [1](#_ENREF_1). In addition to genetic factors, the investigation of behavioral and environmental factors on complex diseases has been the focus of conventional observational epidemiology studies. Although the observational epidemiology studies have made considerable contributions to identifying possible factors affecting complex diseases, many of these observational findings cannot be confirmed by randomized clinical trials (RCTs). For example, the association between Vitamin E intake and coronary heart disease (CHD) risk identified in an observational study cannot be confirmed by a RCT [2](#_ENREF_2). The non-causal associations observed are mainly due to unadjusted confounders [3](#_ENREF_3). Would the success of genetic studies be helpful for studying the causal effect of exposure variables on complex diseases from observational studies? Mendelian randomization (MR), a principle originally due to Katan [4](#_ENREF_4), is to study the exposure-outcome causal relationship from observational studies using genetic variants that affect the disease outcome only through the exposure variable [5](#_ENREF_5).

By the law of independent assortment of Mendel, the inheritance of two different traits is independent, which generally holds with exclusion of linkage disequilibrium (LD). The transmission of alleles from parents to their children is independent, resulting in that alleles are randomly allocated during conception and the genetic variants’ effects on the exposure variable is not subject to confounders [5](#_ENREF_5). At the population level, the associations between genetic factors and exposure variables are generally not confounded, in particular, not confounded by socioeconomic status and behavioral factors [6](#_ENREF_6). In addition, the association between genetic factors and disease risk cannot be due to reverse causality [5](#_ENREF_5). Therefore, genetic variants can be used as a proxy for an exposure variable to study its effect on a disease outcome.

MR study is an application of instrumental variable (IV) analysis by using genetic variants as IVs to study the causal effect of an exposure variable on a disease outcome. IV analysis is commonly used in econometrics to make causal inference from observational studies. As depicted in Figure 1, being valid IVs, genetic variants must satisfy three assumptions: 1. the genetic variants are associated with the exposure variable; 2. the genetic variants are independent of the confounders that confound the association between the exposure variable and the disease outcome; 3. the genetic variants only affect the disease outcome through the exposure variable. The third assumption is also known as the exclusion restriction assumption and indicates the independence between genetic variants and disease outcome given the exposure variable. The three assumptions are sufficient for testing the causal effect of the exposure variable on the disease outcome. At least another assumption, that is, all the associations in Figure 1 are linear and are not subject to interactions, is needed for causal effect size estimation [5](#_ENREF_5).

The disease outcomes in epidemiology studies are often binary, for example, type 2 diabetes (T2D). The most commonly used IV analysis method for a binary outcome is the two-stage residual inclusion (2SRI) [7](#_ENREF_7). Specifically, a linear regression model is fitted in the first stage for the exposure variable using the IV(s) and measured covariates. In the second stage, a nonlinear model is fitted for the binary disease outcome using the exposure variable, measured covariates, and residuals from the first stage linear regression model. The general idea of 2SRI is using the residuals from the first stage regression as a proxy of the unmeasured confounders. 2SRI is also known as the control function method for IV analysis [8](#_ENREF_8) and has been used in MR studies with binary disease outcomes [9](#_ENREF_9). A Wald test using robust standard errors or bootstrapped standard errors is recommended for testing the significance of the causal effect in 2SRI [8](#_ENREF_8).

MR analysis is a very attractive approach for causal inference analysis from observational studies, especially for the exposure variables that are difficult to be studied using RCTs, such as body mass index (BMI). However, MR studies are subject to strong assumptions. Although the first two assumptions are not difficult to satisfy as described above, there may be violations of the exclusion restriction assumption, including pleiotropy, LD, and population stratification [5](#_ENREF_5); [10](#_ENREF_10). The exclusion restriction assumption is often very difficult to validate. Moreover, a single genetic variant, usually a single nucleotide polymorphism (SNP), only explains a small proportion of the total variation in the exposure variable. Hence the genetic variant might turn out to be a weak IV, which can be a challenge in the MR studies, leading to inconclusive and insignificant results. Multiple SNPs or a genetic risk score (GRS), which is a weighted count of effect alleles of multiple SNPs, have been used to alleviate the weak IV problem and increase the statistical power in MR studies [9](#_ENREF_9); [11](#_ENREF_11); [12](#_ENREF_12). In addition, the possible violation of the exclusion restriction assumption can be alleviated by using multiple SNPs in MR studies [5](#_ENREF_5).

Another challenge in MR studies is that the exposure variables of interest are often time-varying (Figure 2), for example, BMI and high-density lipoprotein (HDL). Current MR analyses only use the exposure variable data from some single arbitrary time point, usually at the baseline level [9](#_ENREF_9); [13](#_ENREF_13), although longitudinal data of the time-varying exposures have been collected in many prospective cohort studies, for example, the Framingham Heart Study (FHS) [14](#_ENREF_14). A time-varying exposure may vary substantially over time, and a single measurement is often not adequate in capturing the time-varying information. Furthermore, the exposure variables that change continuously over one’s lifespan might have a cumulative effect on the risk of a disease, e.g., lifelong reduced plasma levels of triglycerides-rich lipoproteins reduce the risk of CHD [15](#_ENREF_15). Davis et al. pointed out that using a single measure of a time-varying exposure could underestimate the relationship between the exposure variable and the outcome variable [16](#_ENREF_16). More importantly, it is the lifetime genetic effect on the exposure that is assumed and estimated in the MR analysis framework, which, however, is unlikely to be obtained with cross-sectional exposure data [17](#_ENREF_17). Given the limitation, how should the longitudinal data of a time-varying exposure variable be incorporated in an MR analysis? Here we propose modeling longitudinal data using functional data analysis methods. Specifically, the underlying individual curve of the time-varying exposure variable is recovered using the functional principal component analysis through conditional expectation (PACE) method, which is designed for modeling irregular and sparse longitudinal data [18](#_ENREF_18). Then MR analysis is performed by assuming that the time-varying exposure variable has a cumulative effect or burden on the risk of the disease. We propose two methods. For the first method, the cumulative burden of the time-varying exposure variable is calculated from the recovered curves and standard MR analysis is then performed using the cumulative burden of the exposure variable. The second method further assumes a time-varying genetic effect on the exposure variable, for which the functional regression models are used. Note that here we focus on testing if a time-varying exposure variable has causal effect on the risk of a disease, rather than effect size estimation. Our simulation studies mimicking real data show that the proposed functional data analysis-based methods incorporating longitudinal data have higher statistical power than the standard MR analysis using a single measurement of the time-varying exposure variable. The proposed method that assumes a time-varying genetic effect has the highest statistical power. We also demonstrated promising performance of the proposed methods by investigating if BMI has causal effect on the risk of T2D and CHD, and if HDL has causal effect on the risk of CHD using the FHS longitudinal data. To the best of our knowledge, this is the first work aiming to incorporate longitudinal data of a time-varying exposure variable in MR analysis to test its causal effect on a binary disease outcome.

## 

## Material and Methods

### Notation

We consider a cohort study for MR analysis. Suppose the cohort study has a total sample size of subjects. Let be the binary disease outcome recorded at time , or 1, for . Let be the longitudinal data vector of the time-varying exposure variable of subject . The time when is recorded is , where for . denotes the vector of multiple SNPs or the GRS of subject . is the vector of observed covariates of subject . Neither of and is time-varying.

### MR analysis using baseline measurement of a time-varying exposure variable

Conventional MR analysis for a time-varying exposure variable is to perform 2SRI using only one measurement among the longitudinal data, usually at the baseline, for example, . Specifically, a linear regression model is fitted for in the first stage: to obtain the fitted residual . Then a logistic regression model is fitted in the second stage: , where is the expected value of . To test the null hypothesis that the time-varying exposure variable has no effect on the disease outcome, i.e., , a Wald test using the robust standard error can be performed [9](#_ENREF_9); [19](#_ENREF_19); [20](#_ENREF_20) . The limitation of the conventional analysis method is that a single measurement is not adequate to represent the level of the exposure variable that changes over time, leading to possible loss of power and spurious results.

### Functional data analysis methods for MR analysis with a time-varying exposure variable

A time-varying exposure variable changes continuously over time. It is intrinsically functional data with longitudinal observations collected at only several time points, as shown in Figure 2. Therefore, we propose to use functional data analysis techniques to incorporate longitudinal data in MR analysis.

*PACE*

Longitudinal data from observational studies are often sparse and collected at irregular time points. In addition, different subjects may have different numbers of observations. Given the characteristics of the longitudinal data, we propose to use the PACE method [18](#_ENREF_18) to recover the underlying curves. The PACE method was developed specifically for modeling sparse longitudinal data with irregular observations by assuming that longitudinal observations of each subject are sampled from an underlying curve with noise and the curves of all the subjects are independent with the same mean function and covariance function [18](#_ENREF_18); [21](#_ENREF_21). In details, let be the mean function and be the covariance function of the collection of curves in a closed time interval , where . Eigendecomposition can be performed to expand the covariance function as , where ’s are nonnegative eigen-values () and ’s are eigen-functions. Then the curve of subject can be expressed as by the Karhunen-Loève theorem [18](#_ENREF_18). is the th functional principal component (FPC) score of subject with mean of 0 and variance of , where . The numerical integration method for FPC score calculation works well for densely observed data, but not for sparse longitudinal data. To solve the problem, Yao et al. [18](#_ENREF_18) introduced additive measurement errors into the model as , where measurement error is assumed to follow the classical measurement error assumption with mean of 0 and variance of [22](#_ENREF_22). For sparse data, the best prediction of is the conditional expectation by assuming that and are jointly normally distributed, where , and [18](#_ENREF_18). When applied to real data, and are estimated by pooling all the observations together. The mean function is estimated using a local linear smoother. The covariance function is estimated by smoothing the sample covariance function using a local linear smoother in the direction of the diagonal and a local quadratic smoother in the direction orthogonal to the diagonal to take into account measurement errors. Eigendecomposition is then performed for after discretization and can be calculated by plugging in parameter estimates from the previous steps. The last step is to recover individual curve using the leading eigen-functions as for . The selection of the number of eigen-functions can be based on the fraction of variance explained (FVE), Akaike information criterion (AIC), or Bayesian information criterion (BIC) [18](#_ENREF_18). The PACE method has been implemented in R package “PACE”.

MR analysis is subject to strong assumptions. For effect estimation, a more strict assumption that all the associations in Figure 1 are linear is required [5](#_ENREF_5). For binary disease outcome, the linear association assumption cannot be satisfied. Therefore, we focus on testing if a time-varying exposure variable has causal effect on a binary disease outcome, not effect size estimation. This is consistent with that the focus of MR analysis is to identify causal risk factors for a disease, not to obtain precise estimation of the effect size [23](#_ENREF_23). We propose two methods for testing the causal effect.

*New Method I: PACE+2SRI*

The first method assumes that the time-varying exposure variable has cumulative burden on the risk of the disease. In detail, the cumulative value of the time-varying exposure variable can be calculated from the recovered curve by integration:

where is the lower bound of the time interval and is the disease incidence time or the time that the follow-up of is censored. Then standard 2SRI can be performed using the cumulative value . A linear regression model is fitted in the first stage:

to obtain the fitted residual . A logistic regression model is then fitted in the second stage:

A Wald test based on the robust standard error is used to test the null hypothesis that the time-varying exposure variable has no effect on the disease risk, i.e., . We denote this method as PACE+2SRI.

*New Method II: PACE+2SFRI*

As gene expression levels change over time, the genetic effect on a time-varying exposure variable changes over time as well. To take this phenomenon into account, we propose the second method for conducting MR analysis for the recovered functional data using functional regression techniques. Specifically, let be the functional data of the time-varying exposure variable with trend removed. A two-stage functional residual inclusion (2SFRI) can be performed. In the first stage, we fit a functional linear model for the time-varying exposure variable as:

,

to obtained the fitted residual . The functional linear model has been implemented in the R package “fda” [24](#_ENREF_24). In the second stage, a functional logistic regression model is fitted to assess the effect of the time-varying exposure variable on the binary disease outcome:

For hypothesis testing purpose, we assume a time-constant effect for both the exposure variable and the fitted residual so that the functional logistic regression model (5) can be simplified to a logistic regression model:

(6)

where , and . To test the null hypothesis that the time-varying exposure variable has no effect on the disease outcome, i.e., , we propose to use a Wald test with the robust standard error. We denote this method as PACE+2SFRI. Of note, the simplified model (6) is in line with the “burden test” in association testing for rare variants [25](#_ENREF_25).

### Simulation Studies

We performed simulation studies to evaluate the performance of proposed functional data analysis-based methods in comparison with the standard MR analysis using only the baseline measurement of a time-varying exposure variable. We considered two simulation set-ups.

*Simulation Set-up I*

We simulated data by assuming that genetic variants have time-varying effect on the exposure variable and the exposure variable has cumulative burden on the disease risk. To mimic the real data, we used the parameter estimates obtained from studying the effect of BMI on the risk of T2D using the GRS as the IV in the FHS data analysis as described in the next section. In detail, we simulated the time-varying exposure of subject as: , where and . was the age that the disease outcome was recorded, which was randomly sampled from the age interval from 51 to 60 years old. , the GRS of subject , was simulated from . Sex was simulated from and then standardized to have mean of 0. represents the unmeasured confounders, simulated from the standard normal distribution. We further let , which was the mean BMI function over time estimated from the PACE procedure. and were the estimated time-varying coefficients of GRS and sex from the first stage of 2SFRI when analyzing the effect of BMI on the risk of T2D using the FHS data. We simulated data using and or to assess the performance of different methods using different IV strength levels. In addition, . The residual was resampled from the BMI residuals estimated from the first stage of 2SFRI. Then the binary disease outcome was simulated from binomial distribution , where . The estimated coefficient of the cumulative burden of BMI was 0.0054 from analyzing its effect on the risk of T2D using the FHS data. We let or 2 to simulate data with different causal effect size. We also simulated data with to check if Type I error rates of the functional data analysis-based methods can be well controlled. To simulate sparse and irregular longitudinal measurements of the time-varying exposure variable, we randomly selected two to five observations at different age points for each subject from the simulated time-varying exposure variable as the observed data. We analyzed each simulated data set using three methods, 2SRI using only the first measurement of the time-varying exposure variable, PACE+2SRI, and PACE+2SFRI. was used as the IV, and sex and were included as observed covariates in all the analysis. The number of FPCs in the PACE procedure was selected based on at least 95% of FVE. We fixed the sample size of each simulated data sets at n=1000. Using the significance level of 0.05, we conducted 2000 replications for empirical Type I error rate evaluation and1000 replications for empirical statistical power evaluation. For different IV strength levels, we calculated the mean *F*-test statistic for testing the association between the GRS and longitudinal exposure data using linear mixed models (LMMs).

*Simulation Set-up II*

To mimic the real data, we simulated the longitudinal data of the time-varying exposure variable and genetic variants by resampling from the FHS data. We further simulated the binary disease outcome using FPC scores instead of assuming a cumulative burden of the exposure variable on the disease risk. Let be the data vector of subject in the FHS data, which was used to analyze the effect of BMI on the risk of T2D. The vector was the longitudinal BMI data collected from seven clinical visits. Many subjects had missing BMI values, meaning that not all the subjects had seven measurements. were the 14 SNPs used for constructing the GRS, and was the time that the disease outcome was recorded. and were the top three FPC scores selected in the PACE procedure to recover the BMI curve. More detailed information on the FHS data is described in the next section. The data vectors in the simulated data sets were resampled with replacement from the observed FHS data vectors We fixed the sample size of each simulated data set at n=1722, the same as the real data. We used parameter estimates from the real data analysis to simulate binary disease outcomes. For statistical power evaluation, the binary disease outcome of subject in a simulated data set was generated from binomial distribution , where . For empirical Type I error rate evaluation, the binary disease outcome of subject in a simulated data set was generated from binomial distribution , where . We analyzed the simulated data sets using either 14 SNPs as IVs or the GRS as a single IV. Sex and were included as covariates in all the analysis. With a significance level of 0.05, we simulated 2000 data sets for empirical Type I error rate evaluation and 1000 data sets for empirical statistical power evaluation.

### Application to the FHS data

To demonstrate the proposed functional data analysis-based methods for testing the causal effect of a time-varying exposure variable on a binary disease outcome, we analyzed the effect of BMI on the risk of T2D and CHD, and the effect of HDL on the risk of CHD using the FHS data. The FHS is a family-based prospective cohort study with subjects from three generations: the Original Cohort, the Offspring Cohort, and the Third Generation Cohort [14](#_ENREF_14). The Offspring Cohort is the largest cohort with both genetic information and longitudinal phenotype measurements collected from seven clinical visits [26](#_ENREF_26). We performed MR analyses using unrelated individuals from the Offspring Cohort. The FHS data was downloaded from NCBI dbGaP. We selected the age interval from 25 to 75 years old to recover the BMI curves and HDL curves from the longitudinal data using the PACE method. We chose such an age interval that at least 50 measurements of an exposure variable were collected at each age point when the data were pooled together to ensure stable estimation of both the mean function and the covariance function in the PACE procedure. The disease outcomes were censored at the last clinic visit that was at or before 75 years old. The subjects included in the analysis had at least two measurements of the exposure variable in the age interval from 25 to 75 years old before the disease incidence or censoring. The sample size for analyzing the effect of BMI on the risk of T2D was 1722 with 171 cases, while those for analyzing the effects of BMI and HDL on the risk of CHD were 1709 with 113 cases and 1669 with 110 cases, respectively. Following the previous MR analyses of the causal effect of BMI on cardiometabolic traits and events [9](#_ENREF_9), we used the 14 selected BMI SNPs identified from large-scale meta-analyses [27](#_ENREF_27). The SNP data were extracted from the FHS Candidate Gene Association Resource (CARe) study. The weights used for the BMI GRS calculation were the same as those in Holmes et al. [9](#_ENREF_9). To analyze the causal effect of HDL on the risk of CHD, we used the 14 HDL SNPs as used in a previous MR study [13](#_ENREF_13). We extracted 2 of the 14 HDL SNPs from the FHS CARe study data and the rest 12 HDL SNPs were imputed from the FHS Affymetrix 500K array using the 1000 Genomes Project haplotypes as the reference panel. We used SHAPEIT for phasing [28](#_ENREF_28) and IMPUTE2 for imputation [29](#_ENREF_29). The HDL GRS was constructed following Voight et al. [13](#_ENREF_13). We conducted MR analyses using both the functional data analysis-based methods and the standard MR method using only the baseline measurement. We used the first measurement of the time-varying exposure variable in the age interval from 25 to 75 years old as the baseline measurement. We used either the GRS as a single IV or the SNPs used in GRS calculation as multiple IVs. We included sex and age when the disease outcome was recorded as covariates.

## Results

### Simulation Studies

In simulation set-up I, we simulated data by assuming that the time-varying exposure variable has a cumulative burden on the disease risk. Table 1 shows the empirical Type I error rates of different analysis methods in the presence of unmeasured confounders. At the significance level of 0.05, all the MR analysis methods were able to control Type I error rates at the nominal level, while direct association analyses using logistic regressions had inflated Type I error rates. Table 2 shows the empirical power of the three MR analysis methods for a time-varying exposure variable, i.e., 2SRI using only the baseline measurement, PACE+2SRI, and PACE+2SFRI, at varying IV strength levels and different cumulative effect sizes of the time-varying exposure variable on the disease risk. As expected, the power increased as either the IV strength or the causal effect size increased. The two functional data analysis-based methods always had higher statistical power than the standard MR analysis method that uses only the baseline measurement in each of the simulated scenario (Table 2).

In simulation set-up II, we simulated the longitudinal measurements of the time-varying exposure variable by resampling from the FHS real data. The disease outcomes were simulated using the FPC scores without assuming a cumulative burden from the exposure variable. Table 3 shows the empirical Type I error rate and power of different MR analysis methods using either the GRS as the IV or 14 SNPs as multiple IVs. All the methods were able to control the Type I error rates at the nominal level. We observe substantial power gain by using the functional data analysis-based methods, especially the PACE+2SFRI method. When comparing MR analysis using the GRS as a single IV with that using the SNPs that were included in the GRS calculation as multiple IVs, all three methods had power increase by approximately 20%. We also performed MR analysis at each visit (see supplemental methods for details) in simulation set-up II. To use the minimum p-value (minP) obtained from analyzing the exposure variable data in each of the seven visits individually, we observe that the Bonferroni correction is necessary to control the Type I error rate (Table S1). The power of the Bonferroni corrected minP method was much lower than that of the baseline method only method, PACE+2SRI, and PACE+2SFRI.

### The FHS Data Analysis

We performed MR analyses of the causal effect of BMI on the risk of T2D and CHD, and the causal effect of HDL on the risk of CHD, using data from the FHS Offspring Cohort. Figure 3 shows the longitudinal BMI data and HDL data that were included in MR analyses. For both BMI and HDL, the individual fluctuation patterns could be very different. While some subjects had relatively stable trajectories over time, others had substantial changes across visits, resulting in that the time-varying information cannot be captured by a single measurement. We used the PACE method to recover the individual BMI curves and HDL curves from the longitudinal data. The smoothed mean BMI function increased slowly over time. The smoothed mean HDL function was almost constant over time (Figure S1). We used the leading three eigen-functions (Figure S2) and their corresponding FPC scores in BMI functional data recovery, which explained 84.7%, 13.6%, and 1.5% of the variation, respectively, with a total of 99.8% variation explained. The first eigen-function shows an almost constant shift from the mean curve. The second eigen-function shows a slowly increase trend over time, while the third one increases from 25 to approximately 50 years old and then decrease. Figure 4 shows the observed longitudinal BMI data and the recovered time-varying underlying curves of four randomly selected subjects. Although the trajectories of different subjects varied, the PACE-recovered curves were able to capture the different patterns. The recovered curves either went through or were adjacent to the longitudinal data points, confirming that majority of the variation in the observed data were well captured. For HDL, we used the leading two eigen-functions (Figure S2) and their corresponding FPC scores in the functional data recovery, which explained 99% and 0.9% of the variation, respectively, with a total of 99.9% variation explained. The first eigen-function shows a very slowly increase trend over time. The second eigen-function shows an increase trend to approximately 55 years old, followed by a decrease. Figure S3 shows the observed longitudinal HDL data and the recovered time-varying underlying curves of four randomly selected subjects. Again, the PACE method was able to recover the time-varying information for subjects with different patterns.

Before performing the MR analyses, we checked the strength of association between the GRS and longitudinal exposure data using a LMM. The *F*-test statistic for the association between the BMI GRS and longitudinal BMI data was 13.25, higher than the weak IV threshold of *F*-test statistic of 11 [11](#_ENREF_11). The *F*-test statistic for the association between the HDL GRS and longitudinal HDL data was 8.94, below the weak IV threshold. Observation analysis of the association between GRS and disease outcomes shows significant association between the BMI GRS and the risk of T2D (p-value of 0.026). The association between the BMI GRS and the risk of CHD was not significant (p-value of 0.793), and neither was the association between the HDL GRS and the risk of CHD (p-value of 0.967).

The MR analysis results of the FHS data are shown in Table 4. When analyzing the causal effect of BMI on the risk of T2D using the GRS as the IV, PACE+2SRI had the smallest p-value of 0.008, more significant than the baseline only analysis (p-value of 0.014), and the PACE+2SFRI analysis (p-value of 0.057). When using the 14 BMI associated SNPs as IVs, MR analysis using the baseline BMI data, PACE+2SRI, and PACE+2SFRI had p-values of 0.061, 0.026, and 0.089, respectively. We did not identify a significant causal effect of either BMI or HDL on the risk of CHD, no matter we used the GRS as the IV or the SNPs as multiple IVs in the MR analyses. When performing observation analysis of the association between BMI and the risk of CHD, the baseline BMI had a significant association, while the cumulative BMI did not. The observation analysis of the association between HDL and the risk of CHD shows a significant association with the cumulative HDL level, but not the baseline HDL level. We also analyzed the effect of BMI on the risk of T2D and CHD, and the effect of HDL on the risk of CHD by individual clinical visits (Table S3, S4, and S5). We only used the exposure data collected between age 25 and 75 years old to be consistent with the data used for functional data analysis-based methods. The sample sizes of different visits were different, but comparable. However, the results are not consistent across clinical visits. For example, in MR analysis of the causal effect of BMI on the risk of T2D using the GRS as the IV (Table S3), the result based on the BMI measurement from the 6th clinical visit was very significant (p-value of 0.002), while the result based on the BMI measurement from the 7th clinical visit was not significant (p-value of 0.097).

## Discussion

In this work, we have proposed novel functional data analysis-based methods for incorporating longitudinal data of a time-varying exposure variable in the MR analysis when the disease outcome is binary. We have shown that the new methods outperformed the current MR analysis, which uses a single measurement at some arbitrary time point.

MR studies have been widely used in recent years aiming to identify causal risk factors from observational studies, especially for the factors that cannot be studied using RCTs, for example, BMI, HDL and alcohol consumption [9](#_ENREF_9); [13](#_ENREF_13); [30](#_ENREF_30). Many risk factors change continuously over one’s lifespan, resulting in that a single measurement cannot capture the time-varying information. However, longitudinal data have not been taken into account in current MR studies, and only one measurement of a time-varying exposure variable, usually at the baseline, is used. The functional data analysis methods we introduce here incorporate longitudinal data in the analysis, and assume that the time-varying exposure variable has a cumulative burden or effect on the risk of developing a disease. We have seen an increase in statistical power in our simulation studies using the functional data analysis-based methods, no matter if the data was simulated with or without the cumulative effect assumption. The PACE+2SFRI method had substantial higher power in simulation set-up II, where the longitudinal data of the exposure variable was resampled from the FHS real data. This may imply that the genetic variants have a time-varying effect on the exposure variable, which leads to a power gain in the statistical analysis that takes into account of time-varying information. In the FHS data analysis, the estimated functional coefficient of the BMI GRS changes over time (Figure 5). The estimated genetic effect on BMI increases slowly from age 25 to 50 years old, followed by a slow decrease.

The GRS is widely accepted in current MR studies to increase IV strength by combining the effect of multiple SNPs [9](#_ENREF_9); [12](#_ENREF_12); [13](#_ENREF_13); [31](#_ENREF_31), where the weights used to construct the GRS are effect estimates from previous large-scale association studies. For example, Holmes et al. [9](#_ENREF_9) constructed the BMI GRS using SNP effect estimates from an association study with 108,912 subjects [27](#_ENREF_27). In our simulation set-up II, we observed a statistical power increase by using the SNPs in the GRS as separate IVs, as compared with using the GRS as a single IV (Table 3). The SNP weights used in the GRS calculation, estimated from 108,912 subjects, may not represent the SNP effect sizes well for the FHS data with 1722 subjects that we used in the simulation studies. This might have lead to the power loss of the GRS-based MR analyses. Extensive simulation studies conducted by Pierce et al. [11](#_ENREF_11) found that using multiple genetic variants as separate IVs may result in weak IVs, while combining them could lead to statistical power loss, which is in agreement with our simulation results. Therefore, we suggest that researchers conduct MR analysis using both IV methods.

In the FHS data analysis, we identified a significant causal effect of BMI on the risk of T2D using the PACE+2SRI method. When using the PACE+2SFRI method, the effect was not significant based on a significance level of 0.05, but the p-values were close to 0.05. We did not identify a significant causal effect of either BMI or HDL on the risk CHD. Our analysis results are generally consistent with previous MR study results based on much larger sample sizes. Holmes et al. showed that BMI had a causal effect on the risk of T2D, but not on the risk of CHD in a MR study with a sample size of 34,538 [9](#_ENREF_9). Voight et al. showed that HDL did not have a causal effect on the incidence of myocardial infarction in a MR study with a sample size of 53,813 [13](#_ENREF_13).

The limitation of this work is that the proposed new methods are only aimed for hypothesis testing, not for causal effect size estimation. Our simulation studies showed that the new methods were able to control the Type I error rates well, confirming that the new methods are valid for testing purpose. However, statistical models that are valid for testing may not provide consistent estimates. Consistent effect size estimation is very challenging in causal inference and is subject to very strong assumptions. Nevertheless, the main focus of MR studies is to identify disease risk factors, rather than effect size estimation [23](#_ENREF_23).

Although MR analysis is subject to strong assumptions, it is a very useful method to identify potential risk factors from observational studies [17](#_ENREF_17). The proposed methods for incorporating longitudinal data of a time-varying exposure will help improve the versatility of MR analysis.

**Acknowledgements**

## References

1. Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187, 367-383.

2. Hooper, L., Ness, A.R., and Smith, G.D. (2001). Antioxidant strategy for cardiovascular disease. The Lancet 357, 1705.

3. Vandenbroucke, J.P. (2004). When are observational studies as credible as randomised trials? The Lancet 363, 1728-1731.

4. Katan, M. (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. The Lancet 327, 507-508.

5. Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Statistics in Medicine 27, 1133-1163.

6. Smith, G.D., Lawlor, D.A., Harbord, R., Timpson, N., Day, I., and Ebrahim, S. (2007). Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. PLoS Medicine 4, e352.

7. Terza, J.V., Basu, A., and Rathouz, P.J. (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. Journal of Health Economics 27, 531-543.

8. Wooldridge, J.M. (2010). Econometric analysis of cross section and panel data.(MIT press).

9. Holmes, M.V., Lange, L.A., Palmer, T., Lanktree, M.B., North, K.E., Almoguera, B., Buxbaum, S., Chandrupatla, H.R., Elbers, C.C., and Guo, Y. (2014). Causal effects of body mass index on cardiometabolic traits and events: a mendelian randomization analysis. The American Journal of Human Genetics 94, 198-208.

10. VanderWeele, T.J., Tchetgen, E.J.T., Cornelis, M., and Kraft, P. (2014). Methodological Challenges in Mendelian Randomization. Epidemiology 25, 427-435.

11. Pierce, B.L., Ahsan, H., and VanderWeele, T.J. (2011). Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. International Journal of Epidemiology 40, 740-752.

12. Palmer, T.M., Lawlor, D.A., Harbord, R.M., Sheehan, N.A., Tobias, J.H., Timpson, N.J., Smith, G.D., and Sterne, J.A. (2012). Using multiple genetic variants as instrumental variables for modifiable risk factors. Statistical Methods in Medical Research 21, 223-242.

13. Voight, B.F., Peloso, G.M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M.K., Hindy, G., Hólm, H., Ding, E.L., and Johnson, T. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. The Lancet 380, 572-580.

14. Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B., Fox, C.S., Larson, M.G., and Murabito, J.M. (2007). The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. American Journal of Epidemiology 165, 1328-1335.

15. The TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute. (2014). Loss-of-function mutations in APOC3, triglycerides, and coronary disease. The New England Journal of Medicine 371, 22-31.

16. Davis, C., Rifkind, B.M., Brenner, H., and Gordon, D.J. (1990). A single cholesterol measurement underestimates the risk of coronary heart disease: an empirical example from the Lipid Research Clinics Mortality Follow-up Study. JAMA 264, 3044-3046.

17. Evans, D.M., and Smith, G.D. (2015). Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. Annual Review of Genomics and Human Genetics.

18. Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association 100, 577-590.

19. Cai, B. (2010). Causal inference with two-stage logistic regression-accuracy, precision, and application. PhD thesis (Philadelphia: University of Pennsylvania).

20. Palmer, T.M., Sterne, J.A., Harbord, R.M., Lawlor, D.A., Sheehan, N.A., Meng, S., Granell, R., Smith, G.D., and Didelez, V. (2011). Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. American Journal of Epidemiology 173, 1392-1403.

21. Müller, H.-G. (2009). Functional modeling of longitudinal data. In: Longitudinal Data Analysis (Handbooks of Modern Statistical Methods).(New York: Wiley).

22. Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). Measurement error in nonlinear models: a modern perspective.(New York: Chapman & Hall).

23. Burgess, S. (2013). Identifying the odds ratio estimated by a two‐stage instrumental variable analysis with a logistic regression model. Statistics in Medicine 32, 4726-4747.

24. Ramsay, J., Hooker, G., and Graves, S. (2009). Functional Data Analysis with R and MATLAB.(New York: Springer).

25. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. The American Journal of Human Genetics 83, 311-321.

26. Cupples, L.A., Heard-Costa, N., Lee, M., and Atwood, L.D. (2009). Genetics analysis workshop 16 problem 2: the Framingham Heart Study data. BMC proceedings 3 (Suppl. 7), S3.

27. Guo, Y., Lanktree, M.B., Taylor, K.C., Hakonarson, H., Lange, L.A., Keating, B.J., Fairfax, B.P., Elbers, C.C., Barnard, J., and Farrall, M. (2013). Gene-centric meta-analyses of 108 912 individuals confirm known body mass index loci and reveal three novel signals. Human Molecular Genetics 22, 184-201.

28. Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. Nature Methods 9, 179-181.

29. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics 44, 955-959.

30. Holmes, M.V., Dale, C.E., Zuccolo, L., Silverwood, R.J., Guo, Y., Ye, Z., Prieto-Merino, D., Dehghan, A., Trompet, S., and Wong, A. (2014). Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. BMJ 349, g4164.

31. Proitsi, P., Lupton, M.K., Velayudhan, L., Newhouse, S., Fogh, I., Tsolaki, M., Daniilidou, M., Pritchard, M., Kloszewska, I., and Soininen, H. (2014). Genetic predisposition to increased blood cholesterol and triglyceride lipid levels and risk of Alzheimer disease: A Mendelian randomization analysis. PLoS Medicine, e1001713.

Table 1. Empirical Type I error rates of different analysis methods in simulation set-up I.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| IV strength | Mean *F-*test statistic  from LMM | MR analysis | | |  | Observation analysis | |
| Baseline 2SRI | PACE+2SRI | PACE  +2SFRI |  | Baseline | PACEa |
|  | 8.34 | 0.053 | 0.056 | 0.056 |  | 0.076 | 0.083 |
|  | 30.29 | 0.053 | 0.056 | 0.057 |  | 0.073 | 0.081 |
|  | 118.02 | 0.053 | 0.054 | 0.052 |  | 0.073 | 0.079 |

a: the cumulative burden of the exposure variable calculated from PACE-recovered curve was tested.

Table 2. Empirical statistical power of different MR analysis methods in simulation set-up I.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| IV strength | Cumulative effect of on |  | MR analysis | | |
|  | Baseline 2SRI | PACE+2SRI | PACE+2SFRI |
|  |  |  | 0.133 | 0.135 | 0.139 |
|  |  | 0.269 | 0.278 | 0.277 |
|  |  |  |  |  |  |
|  |  |  | 0.258 | 0.268 | 0.272 |
|  |  | 0.586 | 0.593 | 0.603 |
|  |  |  |  |  |  |
|  |  |  | 0.588 | 0.596 | 0.601 |
|  |  | 0.980 | 0.986 | 0.986 |

Table 3. Empirical Type I error rates and statistical power of MR analysis methods in simulation set-up II.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| IV |  | MR analysis | | |
|  | Baseline 2SRI | PACE+2SRI | PACE+2SFRI |
| *Empirical Type I error rate* | | | | |
| GRS |  | 0.041 | 0.041 | 0.046 |
| 14 SNPs |  | 0.059 | 0.056 | 0.053 |
| *Empirical statistical power* | | | | |
| GRS |  | 0.188 | 0.206 | 0.259 |
| 14 SNPs |  | 0.380 | 0.417 | 0.450 |

Table 4. Analysis of the causal effect of BMI on the risk of T2D and CHD, and the effect of HDL on the risk of CHD using the FHS data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Exposure | Disease | MR analysis p-value | | | |  | Observation analysis p-value | |
| IV | Baseline 2SRI | PACE+2SRI | PACE +2SFRI | Baseline | PACEa |
| BMI | T2D | GRS | 0.014 | 0.008 | 0.057 |  | < 2E-16 | < 2E-16 |
| 14 SNPs | 0.061 | 0.026 | 0.089 |  |
| BMI | CHD | GRS | 0.396 | 0.397 | 0.510 |  | 0.005 | 0.708 |
| 14 SNPs | 0.438 | 0.467 | 0.532 |  |
| HDL | CHD | GRS | 0.518 | 0.531 | 0.239 |  | 0.114 | 0.005 |
| 14 SNPs | 0.783 | 0.790 | 0.489 |  |

a: the cumulative burden of the exposure variable calculated from PACE-recovered curve was tested.

Figure 1. Directed acyclic graph of MR analysis assumptions.

Figure 2. Directed acyclic graph of MR analysis with a time-varying exposure variable. Blue points represent observed longitudinal data. The black line represents unobserved level that changes continuously over time.



Figure 3. The observed FHS longitudinal data. For the plot on the left and the plot in the middle, each line represents the longitudinal BMI data of a subject included in MR analysis. For the plot on the right, each line represents the longitudinal HDL data of a subject included in MR analysis.



1141

Figure 4. PACE-predicted vs. observed BMI data. Each plot shows the BMI data of a randomly selected subject.



Figure 5. The estimated functional coefficient of the BMI GRS from the first stage of 2SFRI. The dotted lines indicate the point-wise 95% confidence limits



## Supplemental Data

***Mendelian Randomization (MR) Analysis by Individual Clinical Visits.***

For cohort studies with a fixed number of clinical visits, an alternative way of conducting MR analysis for the longitudinal exposure data is to perform standard MR analysis by using the exposure data collected from each visit and then calculate the minimum p-value (minP). For example, the Offspring Cohort of the FHS study had seven clinical visits. We can conduct the standard MR analysis for each visit and then use the minP of the seven analyses as the final result. Following the notation in the main text, we let , meaning that all the subjects had the same number of measurements for the exposure variable. For , 2SRI is conducted. Specifically, a linear regression model is fitted for in the first stage: to obtain the fitted residual . Then a logistic regression model is fitted in the second stage: , where is the expected value of . We can obtain the p-value from testing the null hypothesis using a Wald test with the robust standard error, for . Then minP is compared with the Bonferroni corrected significance level, for example, . The limitations of the minP method are: first, not all the subjects have data collected for each clinical visit, leading to varying sample size of each analysis, and second, the Bonferroni correction may lead to power loss.

**Figure S1**. The smoothed mean BMI function and mean HDL function by the PACE procedure.



**Figure S2**. The corresponding eigen-functions of the top FPC scores estimated for FHS longitudinal BMI data and HDL data by the PACE procedure. The plot on the left shows the three estimated BMI eigen-functions corresponding to the top three FPC scores. The plot on the right shows the two estimated HDL eigen-functions corresponding to the top two FPC scores.



**Figure S3**. PACE-predicted vs. observed HDL data. Each plot shows the HDL data of a randomly selected subject.



**Table S1.** Empirical Type I error rates of MR analysis by individual clinical visits in simulation set-up II.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| IV | Visit1 | Visit2 | Visit3 | Visit4 | Visit5 | Visit6 | Visit7 | Uncorrected minP | Bonferroni corrected minP |
| GRS | 0.041 | 0.044 | 0.047 | 0.044 | 0.043 | 0.047 | 0.042 | 0.104 | 0.018 |
| 14 SNPs | 0.058 | 0.050 | 0.049 | 0.052 | 0.058 | 0.059 | 0.060 | 0.161 | 0.036 |

**Table S2.** Empirical statistical power of MR analysis by individual clinical visits in simulation set-up II.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| IV | Visit1 | Visit2 | Visit3 | Visit4 | Visit5 | Visit6 | Visit7 | Bonferroni corrected minP |
| GRS | 0.155 | 0.146 | 0.139 | 0.164 | 0.188 | 0.173 | 0.151 | 0.083 |
| 14 SNPs | 0.329 | 0.266 | 0.323 | 0.359 | 0.369 | 0.329 | 0.338 | 0.283 |

**Table S3.** Analysis of the causal effect of BMI on the risk of T2D by individual clinical visits using the FHS data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Clinical visit | Sample size | MR analysis p-value | | Observation analysis p-value |
| GRS | 14SNPs |
| 1 | 1515 | 0.023 | 0.167 | 9.58E-22 |
| 2 | 1444 | 0.004 | 0.015 | 6.55E-24 |
| 3 | 1489 | 0.004 | 0.041 | 2.67E-20 |
| 4 | 1591 | 0.017 | 0.012 | 8.37E-22 |
| 5 | 1583 | 0.009 | 0.002 | 1.76E-20 |
| 6 | 1505 | 0.002 | 0.108 | 3.38E-14 |
| 7 | 1440 | 0.097 | 0.390 | 1.75E-08 |
| Bonferroni corrected minP |  | 0.014 | 0.014 | 4.59E-23 |

**Table S4.** Analysis of the causal effect of BMI on the risk of CHD by individual clinical visits using the FHS data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Clinical visit | Sample size | MR analysis p-value | | Observation analysis p-value |
| GRS | 14SNPs |
| 1 | 1503 | 0.399 | 0.448 | 0.007 |
| 2 | 1431 | 0.412 | 0.290 | 0.005 |
| 3 | 1471 | 0.390 | 0.365 | 0.030 |
| 4 | 1574 | 0.281 | 0.162 | 0.111 |
| 5 | 1563 | 0.298 | 0.145 | 0.264 |
| 6 | 1491 | 0.318 | 0.103 | 0.777 |
| 7 | 1454 | 0.132 | 0.241 | 0.734 |
| Bonferroni corrected minP |  | 0.924 | 0.721 | 0.035 |

**Table S5.** Analysis of the causal effect of HDL on the risk of CHD by individual clinical visits using the FHS data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Clinical visit | Sample size | MR analysis p-value | | Observation analysis p-value |
| GRS | 14SNPs |
| 1 | 1438 | 0.413 | 0.894 | 0.340 |
| 2 | 1366 | 0.493 | 0.753 | 0.015 |
| 3 | 1398 | 0.373 | 0.462 | 0.020 |
| 4 | 1506 | 0.656 | 0.870 | 0.028 |
| 5 | 1525 | 0.574 | 0.889 | 0.004 |
| 6 | 1450 | 0.682 | 0.782 | 0.192 |
| 7 | 1416 | 0.317 | 0.585 | 0.965 |
| Bonferroni corrected minP |  | 1 | 1 | 0.028 |

# Chapter 5: Conclusion and Future Directions

Complex diseases are affected by many risk factors, including genetic factors, environmental factors, and their interactions. Risk factor identification for complex diseases is challenge and remains the focus of current epidemiology studies. The first part of this dissertation focuses on statistical method development for vQTL mapping using family data, which provides a new tool for identifying genetic factors, gene-gene and gene-environment interactions for complex diseases. The second part of this dissertation focuses on Mendelian randomization analysis of a time-varying exposure variable. Our newly proposed methods for Mendelian randomization analysis will help study time-varying environmental risk factors of complex diseases, especially the risk factors that cannot be studied using RCTs.

vQTL is a new class of genetic variants that can be used as candidates for identifying gene-gene and gene-environment interactions ([Paré et al., 2010](#_ENREF_21); [Struchalin et al., 2010](#_ENREF_30)). Several statistical methods have been developed or suggested for vQTL identification for uncorrelated data ([Cao et al., 2014](#_ENREF_4); [Paré et al., 2010](#_ENREF_21); [Rönnegård and Valdar, 2011](#_ENREF_23); [Shen et al., 2012](#_ENREF_24); [Struchalin et al., 2012](#_ENREF_29)). However, there were no statistical methods designed specifically for vQTL identification using family-based data. In Chapter 2, we described famLRTs in a LMM framework for family data based vQTL identification. The famLRT test statistic follows a chi-squared distribution when the phenotype residuals are normally distributed. For the phenotypes whose residuals are not normally distributed, we proposed parametric bootstrap-based resampling method for p-value calculation. Simulation studies show that the famLRTs were able to control Type I error rates at the nominal level and had good statistical power. We demonstrated the famLRTs using FHS data for BMI vQTL identification.

The famLRTs filled a gap for family-based vQLT study. One limitation of the famLRTs is that parametric bootstrap is needed for p-value calculation when the phenotype residuals are not normally distributed, which is often the real case. The calculation could be computationally intensive if the significance level is low. Another limitation is that the BLUP of family random effect needs to be removed before parametric bootstrap, due to the model complexity introduced by unique family structures. This may lead to statistical power loss. Future work can aim to develop alternative computationally efficient modeling and testing approaches for family data-based vQLT identification, which can also achieve high statistical power. Another way to overcome the limitations is to find an analytical solution for famLRT p-value calculation when the phenotype residuals are not normally distributed.

Mendelian randomization analysis is a method aiming to make causal inference of an exposure variable’s effect on a phenotype or a disease outcome from observational epidemiology studies ([Lawlor et al., 2008](#_ENREF_17)). It is an application of IV analysis and uses genetic variants as IVs. Many exposure variables of interest are time-varying, whose level changes over one’s lifespan. However, current practice of Mendelian randomization analysis does not take the time-varying information into account and only uses a single measurement of the exposure variable. We proposed functional data analysis-based methods to recover the underlying individual trajectories of the time-varying exposure variable from sparse and irregular longitudinal data. Then the recovered functional data is used to analyze the effect of the time-varying exposure variable. We proposed different Mendelian randomization analysis methods of the recovered functional exposure data for different types of outcome variables. The manuscript on a continuous phenotype outcome variable is in Chapter 3. The manuscript on a binary disease outcome variable is in Chapter 4. For both types of outcome variables, we focused on testing if a time-varying exposure variable has a causal effect on an outcome variable. Simulation studies show that the functional data analysis-based methods for both continuous outcomes and binary outcomes outperformed their corresponding standard Mendelian randomization analysis methods. Incorporating longitudinal data of a time-varying exposure variable helped increase statistical power of Mendelian randomization analysis. We applied the functional data analysis-based Mendelian randomization analysis methods to the FHS data.

To the best of our knowledge, no statistical methods had been proposed specifically for Mendelian randomization analysis of a time-varying exposure variable before this work. We attempted to provide a statistical solution for this challenge in current Mendelian randomization studies. We have shown that the proposed functional data analysis-based methods had advantage in the sense of statistical testing, when compared with standard Mendelian randomization analysis methods. One limitation of this work is that we only focused on statistical testing. We did not investigate the properties of the effect estimators or provide asymptotic properties from the statistical theory perspective.

Given the current progress of Mendelian randomization analysis, I would suggest the following future directions of Mendelian randomization analysis of a time-varying exposure: 1. Simultaneous Mendelian randomization analysis for multiple time-varying exposure variables that are correlated or are affected by overlapping genetic variants; 2. Extending the number of genetic variants from multiple to a much larger number as in current polygenic studies; which will significantly increase the explained variation of the exposure variable by genetic variants and lead to power increase of Mendelian randomization analysis; 3. Modeling the effect of a time-varying exposure variable on a binary disease outcome using survival analysis techniques; 4. Incorporating family structures that Mendelian randomization analysis can be extended to family data; 5. Investigating the asymptotic properties of causal effect estimates of a time-varying exposure variable.

Both the methodological development and application of Mendelian randomization analysis have advanced dramatically in the past decade. Mendelian randomization analysis had been extended from using one genetic variant to using multiple genetic variants (Palmer et al., 2011), and from investigating one exposure variable to investigating multiple exposure variables simultaneously (Burgess and Thompson, 2015). More novel applications and extensions of Mendelian randomization analysis have been suggested, including using Mendelian randomization analysis to inform drug development, using Mendelian randomization analysis to mine the phenome, network Mendelian randomization analysis, and even hypothesis-free causality (Evans and Smith, 2015). With the rapid advancement of the –omics technologies and fast increasing volumn of biological data being generated, the application and potential value of Mendelian randomization analysis is unlimited.

# References

Aulchenko, Y.S., De Koning, D.-J., and Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics *177*, 577-585.

Balding, D.J. (2009). The advance of Bayesian methods for genetic association analysis. Presented at the 59th Annual Meeting of The American Society of Human Genetics *November 22, 2009, Honolulu, HI.*

Bound, J., Jaeger, D.A., and Baker, R.M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. Journal of the American Statistical Association *90*, 443-450.

Burgess, S., & Thompson, S. G. (2015). Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. American Journal of Epidemiology *181(4),* 251-260.

Cao, Y., Wei, P., Bailey, M., Kauwe, J.S., and Maxwell, T.J. (2014). A versatile omnibus test for detecting mean and variance heterogeneity. Genetic Epidemiology *38*, 51-59.

Cardon, L.R., and Palmer, L.J. (2003). Population stratification and spurious allelic association. The Lancet *361*, 598-604.

Chen, H., Meigs, J.B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. Genetic Epidemiology *37*, 196-204.

Craig, J. (2008). Complex diseases: Research and applications. Nature Education *1*, 184.

Davis, C., et al. (1990) A single cholesterol measurement underestimates the risk of coronary heart disease: an empirical example from the Lipid Research Clinics Mortality Follow-up Study. JAMA *264*, 3044-3046.

Doll, R., Peto, R., Boreham, J., and Sutherland, I. (2004). Mortality in relation to smoking: 50 years' observations on male British doctors. British Medical Journal *328*, 1519.

Doll, R., Peto, R., Boreham, J., and Sutherland, I. (2005). Mortality from cancer in relation to smoking: 50 years observations on British doctors. British Journal of Cancer *92*, 426-429.

Evans, D. M., & Smith, G. D. (2015). Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. Annual Review of Genomics and Human Genetics *First posted online on April 22, 2015*.

Holmes, M.V., Lange, L.A., Palmer, T., Lanktree, M.B., North, K.E., Almoguera, B., Buxbaum, S., Chandrupatla, H.R., Elbers, C.C., and Guo, Y. (2014). Causal effects of body mass index on cardiometabolic traits and events: a mendelian randomization analysis. The American Journal of Human Genetics *94*, 198-208.

Hooper, L., Ness, A.R., and Smith, G.D. (2001). Antioxidant strategy for cardiovascular disease. The Lancet *357*, 1705.

Jimenez-Gomez, J.M., Corwin, J.A., Joseph, B., Maloof, J.N., and Kliebenstein, D.J. (2011). Genomic analysis of QTLs and genes altering natural variation in stochastic noise. PLoS Genetics *7*, e1002295.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nature Genetics *42*, 348-354.

Katan, M. (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. The Lancet *327*, 507-508.

Laird, N.M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. Nature Review Genetics *7*, 385-394.

Lawlor, D.A., Harbord, R.M., Sterne, J.A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Statistics In Medicine *27*, 1133-1163.

Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics *95*, 5-23.

Liu, C.T., Monda, K.L., Taylor, K.C., Lange, L., Demerath, E.W., Palmas, W., Wojczynski, M.K., Ellis, J.C., Vitolins, M.Z., Liu, S.*, et al.* (2013). Genome-wide association of body fat distribution in African ancestry populations suggests new loci. PLoS Genetics *9*, e1003681.

Oualkacha, K., Dastani, Z., Li, R., Cingolani, P.E., Spector, T.D., Hammond, C.J., Richards, J.B., Ciampi, A., and Greenwood, C.M. (2013). Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. Genetic Epidemiology *37*, 366-376.

Palmer, T. M., Lawlor, D. A., Harbord, R. M., Sheehan, N. A., Tobias, J. H., Timpson, N. J., ... & Sterne, J. A. (2012). Using multiple genetic variants as instrumental variables for modifiable risk factors. Statistical Methods in Medical Research *21(3)*, 223-242.

Paré, G., Cook, N.R., Ridker, P.M., and Chasman, D.I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. PLoS Genetics *6*, e1000981.

Pierce, B.L., Ahsan, H., and VanderWeele, T.J. (2011). Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. International Journal of Epidemiology *40*, 740-752.

Rönnegård, L., and Valdar, W. (2011). Detecting major genetic loci controlling phenotypic variability in experimental crosses. Genetics *188*, 435-447.

Shen, X., Pettersson, M., Rönnegård, L., and Carlborg, Ö. (2012). Inheritance beyond plain heritability: variance-controlling genes in Arabidopsis thaliana. PLoS Genetics *8*, e1002839.

Smith, G.D., and Ebrahim, S. (2003). ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology *32*, 1-22.

Smith, G.D., Lawlor, D.A., Harbord, R., Timpson, N., Day, I., and Ebrahim, S. (2007). Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. PLoS Medicine *4*, e352.

Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B., Fox, C.S., Larson, M.G., and Murabito, J.M. (2007). The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. American Journal of Epidemiology *165*, 1328-1335.

Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. Genetics *187*, 367-383.

Struchalin, M.V., Amin, N., Eilers, P.H., van Duijn, C.M., and Aulchenko, Y.S. (2012). An R package "VariABEL" for genome-wide searching of potentially interacting loci by testing genotypic variance heterogeneity. BMC Genetics *13*, 4.

Struchalin, M.V., Dehghan, A., Witteman, J.C.M., van Duijn, C., and Aulchenko, Y.S. (2010). Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. BMC Genetics *11*, 92.

Sun, W. (2012). A statistical framework for eQTL mapping using RNA-seq data. Biometrics *68*, 1-11.

Terza, J.V., Basu, A., and Rathouz, P.J. (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. Journal of Health Economics *27*, 531-543.

Vandenbroucke, J.P. (2004). When are observational studies as credible as randomised trials? The Lancet *363*, 1728-1731.

VanderWeele, T.J., Tchetgen, E.J.T., Cornelis, M., and Kraft, P. (2014). Methodological Challenges in Mendelian Randomization. Epidemiology *25*, 427-435.

Voight, B.F., Peloso, G.M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M.K., Hindy, G., Hólm, H., Ding, E.L., and Johnson, T. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. The Lancet *380*, 572-580.

Wei, W.-H., Hemani, G., and Haley, C.S. (2014). Detecting epistasis in human complex traits. Nature Reviews Genetics.

Wooldridge, J. (2012). Introductory econometrics: A modern approach (Cengage Learning).

Wooldridge, J.M. (2010). Econometric analysis of cross section and panel data (MIT Press).

Wu, C., Kraft, P., Zhai, K., Chang, J., Wang, Z., Li, Y., Hu, Z., He, Z., Jia, W., and Abnet, C.C. (2012). Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. Nature Genetics *44*, 1090-1097.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., and Holland, J.B. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics *38*, 203-208.

Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences *109*, 1193-1198.